DISSERTATION PROSPECTUS: COMPUTATIONAL NATURAL LANGUAGE INFERENCE: ROBUST AND INTERPRETABLE QUESTION ANSWERING

by

Rebecca Reynolds Sharp

A Dissertation Submitted to the Faculty of the

DEPARTMENT OF LINGUISTICS

In Partial Fulfillment of the Requirements

For the Degree of

DOCTOR OF PHILOSOPHY

In the Graduate College

THE UNIVERSITY OF ARIZONA

2017

CHAPTER 1

INTRODUCTION

One of the primary goals for natural language processing is the development of tools that are useful for humans both at work and at home. The classic example of one of these tools would be a search engine able to take a query and return a relevant and informative result. In order for this application to be truly useful, though, it needs to operate over natural language queries and natural language results. While certain queries and results can be anticipated using statistics, in order to handle a new query, the system must be able to *infer* what the desired result is from the query. This natural language inference is critical to search as well as many other types of natural language processing tasks including question answering.

Question answering (QA), i.e., finding short answers to natural language questions, is one of the most important but challenging tasks on the road towards natural language understanding Etzioni (2011). As mentioned above, a QA system first faces the challenge that both the question and any potential answer sources are in the form of natural language, which inherently contains a large degree of lexical, syntactic, and even dialectal variation. Additionally, 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*). This requires QA models to draw upon more complex methods to bridge this "lexical chasm" Berger et al. (2000), i.e., to infer what the answer should be. 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.

1.1 Robustness

While first order logic methods are attractive for their formality, it is this same formality that renders them too brittle to be of much use outside of small, toy domains. In order to be used, these methods require parsing natural language into formal meaning representations (a task which is very difficult given the previously mentioned language variations) and then performing inference over these representations using logic rules and axioms, again made more difficult by the open-domain (i.e., questions can be about anything and of any level of complexity) and the fact that there are never enough rules and axioms to encode everything you need to know about the world and language. Here we focus on *approximating* this formal logic inference using natural language and machine learning instead of formal representations and hand-crafted rules. While we lose some of the ability to *prove* an answer, we gain much-needed robustness to both language and domain.

Many of these shallower methods that are based on machine learning (rather than pre-written rules) require a large amount of supervised training data, that is, data for which the label is known. In the context of QA, these are questions for which we know the *true* answer(s)[1]. In many domains, however, this requirement is expensive as large-scale, high-quality data is hard to find. Therefore, in this work, we emphasize methods which use semi- and distant supervision[2], rather than

[1]The true, or gold, answer to a question can be debatable, and there are many situations in which there can be more than one correct answer. Typically, an answer is considered correct if a human expert labels it as such, but again, this is open to interpretation.

[2]With semi-supervision, you begin with a small amount of hand-labeled data and learn to label new examples automatically, expanding your pool of labeled examples. With distant supervision, the labeling is done automatically from the start, e.g., by aligning a database of known facts with free text instances containing portions of those facts. For example, if you have a fact triple such as *(Jane_Smith, president_of, College_X)*, you can find all sentences that contain those three elements and count them all as correct instances of the relation *president_of*. This type of approach is able to scale to large data-sources, but is very prone to noise (e.g., not all sentences truly demonstrate

needing a large amount of hand-generated training data. This emphasis allows these methods to be applicable across a wider variety of domains.

Another potential problem associated with these shallower, machine learning-based methods is that with the relaxation of the formality, the explainability may also be lost, and with it the ability to understand *why* the model produces the answers it does. This model understanding is critical for being able to discern sources of model errors and fix them to attain better performance or generalization. To address this, we include models (Chapters 5 and 6, see also Section 1.2) that use a human-readable intermediate output generated by the model to provide an explanation for the inference performed by the model.

1.1.1 Aligning questions and answers

Berger et al. (2000) proposed that the "lexical chasm", or lack of lexical overlap that often exists between questions and answers, might be partially bridged by making use of techniques popular in machine translation, the task of automatically translating text from one language to another language. At a high level, in typical machine translation, a model is given a sentence in one language aligned with its translation in another language. The model keeps track of statistics such as how frequently words in each language co-occur in the sentence-translation pairs and uses these statistics to generate word-alignment scores. Berger et al. suggested repurposing these statistical machine translation models for QA. Instead of translating text from one language to another, these *monolingual* alignment models (i.e., models that align words within the same language) learn to translate from question to answer[3]. Given the previously mentioned example question, *What should we eat for breakfast? – Zoe's Diner has good pancakes*, these models learn common asso-

the *president_of* relation).

[3]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).

ciations 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 Section **??** for more discussion), they have expensive training data requirements, requiring a large set of aligned in-domain question-answer pairs for training. In most domains these pairs are expensive to generate, and one of the current methodological challenges in QA is locating or building high-quality QA pairs for training and testing. Even large open-domain international evaluations and workshops such as the Text REtrieval Conference (TREC)[4] and the Cross Language Evaluation Forum (CLEF),[5] are often limited to sets of a few hundred factoid questions, many of which are highly related. As a result, for open domain QA one often makes use of Community Question Answering data from websites such as Yahoo! Answers or Stack Overflow, which offer tens of thousands of questions, but of highly variable quality. 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)[6] in Chapter 3 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. Specifically we align the head and the dependent of each discourse relation in lieu of a question and answer. This method can be

[4] `http://trec.nist.gov`

[5] `http://www.clef-initiative.eu`

[6] Discourse parsing involves the segmentation of text into adjacent spans which are then recursively joined and assigned both a direction (which span is the head, and which is the dependent) and a label (e.g., *elaboration*, *contrast*, *attribution*, etc).

thought of as a form of distant supervision, as we don't have true labels for which free-text pairs are truly high-quality proxies for question-answer pairs. We evaluate our methods on two corpora, generating alignment models for an open-domain community QA task using Gigaword[7], and for a biology-domain QA task using a biology textbook.

1.1.2 Customizing approaches to a specific question type

This alignment approach for QA can be considered as falling into a larger group of approaches which prefer answers that are closely related to the question, where the relatedness is determined by the associations of the alignment model (e.g., as briefly referred to in Section 1.1.1) or by associations provided by other lexical semantic models such as word embeddings (Yih et al., 2013; Jansen et al., 2014; Fried et al., 2015). While appealing for its robustness to natural language variation, this one-size-fits-all category of approaches does not take into account the wide range of distinct question types that can appear in any given question set (see, e.g., Table 1.2), and that are best addressed individually (Chu-Carroll et al., 2004; Ferrucci et al., 2010; Clark et al., 2013a) for a specific question set.

Given the variety of question types, we suggest that a better approach is to look for answers that are related to the question *through the appropriate relation*, e.g., a causal question, should have a cause-effect relation with its answer. For example, to score a candidate answer for a question such as *What do hurricanes cause?*, while general purpose word associations would give a high score to a sentences containing near associates such as *tornadoes* and *tropical storms*, we suggest that a better approach would be using a set of relation-specific word associations that would provide high scores for words that are in a cause-effect relation with *hurricane*, such as *flooding* or *damage*. Adopting this view, and working with embeddings as a mechanism for assessing relationship, in Chapter 4 we address a key question: how

[7]LDC catalog number LDC2012T21

do we train and use task-specific embeddings cost-effectively? Using causality as a use case, we answer this question with a framework for producing causal word embeddings, i.e., a set of word embeddings which encode causality rather than similarity, with minimal supervision, and a demonstration that such task-specific embeddings significantly benefit causal QA.

1.2 Interpretability

Developing interpretable machine learning models, that is, models where a human user can *understand* what the model is learning, is considered by many to be crucial for ensuring usability and accelerating progress (Craven and Shavlik, 1996; Kim et al., 2015; Letham et al., 2015; Ribeiro et al., 2016). For many applications of question answering (QA), simply providing an answer (or even an answer alongside a list of word-associations and their corresponding model weights, as with the approaches briefly described in Section 1.1) is not sufficient. A complete approach must be interpretable, i.e., able to *explain* why an answer is correct. For example, in the medical domain, a user would not trust a system that recommends invasive procedures without giving a justification as to why (e.g., "Smith (2005) found procedure X healed 90% of patients with heart disease who also had secondary pulmonary complications"). A QA tool is clearly more useful when its human user can identify both when it functions correctly, and when it delivers an incorrect or misleading result – especially in situations where incorrect results carry a high cost.

One approach to interpreting complex models is to make use of human-interpretable information generated by the model to gain insight into what the model is learning. We follow the intuition of Lei et al. (2016), whose two-component network first generates text spans from an input document, and then uses these text spans to make predictions. Lei et al. utilize these intermediate text spans to infer the model's preferences. By learning these intermediate representations end-to-end with a downstream task (i.e., question answering or another task that the user is

Question:	Which of these is a response to an internal stimulus?
(A)	A sunflower turns to face the rising sun.
(B)	A cucumber tendril wraps around a wire.
(C)	A pine tree knocked sideways in a landslide grows upward in a bend.
(**D**)	**Guard cells of a tomato plant leaf close when there is little water in the roots .**
Justification:	Plants rely on hormones to send signals within the plant in order to respond to internal stimuli such as a lack of water or nutrients.

Table 1.1: Example of an 8th grade science question with a justification for the correct answer. Note the lack of direct lexical overlap present between the justification and the correct answer, demonstrating the difficulty of the task of finding justifications using traditional distant supervision methods.

interested in), they are optimized to correlate with what the model learns is discriminatory for the task, and they can be evaluated against what a human would consider to be important.

1.2.1 Multiple choice science questions as a proving ground

In Chapters 5 and 6, we apply this general framework for model interpretability to QA, and in particular to answering multiple-choice science exam questions (Clark, 2015). An example science exam question is provided in Table 1.1 to demonstrate the role of the justification in explaining the correct answer choice as well as to illustrate the fact that retrieving such a justification is non-trivial due to lack of lexical overlap.

In addition to the general difficulties of QA previously described, this particular QA domain is challenging as approximately 70% of science exam questions have been shown to require complex forms of inference to solve (Clark et al., 2013b; Jansen et al., 2016). In Table 1.2 we provide example questions from each of the three main categories of questions from Clark et al. (2013b), where the categories are based on the methods likely required to answer them correctly. Not only do the majority

Category	Example	
Retrieval (35%)	Q: The movement of soil by wind or water is called:	
	(A) condensation	(B) evaporation
	(C) erosion	(D) friction
General Inference (39%)	Q: Which example describes an organism taking in nutrients?	
	(A) A dog burying a bone	(B) A girl eating an apple
	(C) An insect crawling on a leaf	(D) A boy planting tomatoes in the garden
Model-based Inference (26%)	Q: When a baby shakes a rattle, it makes a noise. Which form of energy was changed to sound energy?	
	(A) electrical	(B) light
	(C) mechanical	(D) heat

Table 1.2: Categories of questions and their relative frequencies as identified by Clark et al. (2013b). Retrieval-based questions (including *is–a*, dictionary definition, and property identification questions) tend to be answerable using information retrieval methods over structured knowledge bases, including taxonomies and dictionaries. More complex general inference questions make use of either simple inference rules that apply to a particular situation, a knowledge of causality, or a knowledge of simple processes (such as *solids melt when heated*). Difficult model-based reasoning questions require a domain-specific model of how a process works, like how gravity causes planets to orbit stars, in order to be correctly answered. Note here that we do not include diagram questions, as they require specialized spatial reasoning that is beyond the scope of this work.

of questions require some form of inference to solve, but there are few structured knowledge bases to support this inference, and also commonly incorrect answers that are high semantic associates of either the question or correct answer are included to "lure" students (or automated methods) away from the correct response.

1.2.2 Reranking justifications with weak supervision

Within the domain of multiple-choice science QA, we propose two approaches that reframe QA from the task of scoring (or reranking) answers to a process of *generating and evaluating justifications* for why a particular answer candidate is correct.

Each of these approaches learn to both select and explain answers, when the only supervision available is for which answer is correct (but not how to explain it). Intuitively, our approaches choose the justifications that provide the most help towards ranking the correct answers higher than incorrect ones. More formally, our approaches alternate between using the current model to choose the highest scoring justifications for answers, and optimizing the answer ranking model given these justifications. Thus, for both of these approaches, for each question and candidate answer we gather (or create) a pool of potential justifications. Then we allow the model to learn how to rerank these justifications such that the highest-scoring justification for the correct answer is better than the highest-scoring justification for any of the incorrect answers. Crucially, for both approaches, these reranked texts serve as our human-readable answer justifications, and by examining them, we gain insight into what the model learned was useful for the QA task.

The first approach (Chapter 5) uses aggregation of information from several sources along with structured representations of the text to create justifications of the answer choices. In particular, we aggregate multiple sentences into hierarchical graph structures (called text aggregation graphs, Section 5.4) that capture both intrasentence syntactic structures and intersentence lexical overlaps. Further, we model whether the intersentence lexical overlap is between contextually relevant keywords critical to the justification, or other words which may or may not be relevant.

The second (Chapter 6) uses a shallower approach without aggregation or structured representations. In this shallow approach, we consider only justifications which are single sentences from a corpus and we replace the graph representation of sentences with embeddings and a small set of explicit features (that model lexical overlap, length, etc). These changes allow us to operate over larger text resources.

Despite the differences between these two approaches, however, each allows us to address the challenge of question answering while prioritizing interpretability of

the model.

1.3 Contributions

With this work we tackle the challenging task of question answering, seeking methods which balance the (sometimes conflicting) demands of robustness and interpretability. We address these challenges with four approaches. In particular, the specific contributions of this work are:

1.3.1 Contribution 1: Using discourse structures to generate artificially aligned pairs for training question answering models

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. To this end, 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) and 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. The proposed methods are evaluated on two corpora, including a low-resource domain where training data is expensive (biology). 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.

1.3.2 Contribution 2: Generating customized, task-specific word embeddings based on the question type

We propose a methodology for generating causal word-embeddings (encoding causality rather than similarity) cost-effectively by bootstrapping cause-effect pairs ex-

tracted from free text using a small set of seed patterns, e.g., *X causes Y*. We then train dedicated causal word embedding (as well as two other distributional similarity) models over this data. Levy and Goldberg (2014) have modified the algorithm ofMikolov et al. (2013b) to use an arbitrary, rather than linear, context. Here we make this context *task-specific*, i.e., the context of a cause is its effect. That is, for a word like *hurricane* in the sentence *The hurricane caused extensive flooding and damage*, we replace the standard sliding window context (*the, caused*, and *extensive*) with the words which are the effects: *flooding* and *damage*. Further, to mitigate sparsity and noise, our models are bidirectional, and noise-aware (by incorporating the likelihood of noise in the training process). We implement a QA system that uses these causal embeddings to answer questions and demonstrate that they significantly improve performance over a strong baseline. Further, we show that causal embeddings encode complementary information to the standard (or *vanilla*) word-embeddings, even when trained from the same knowledge resources. We also analyze direct vs. indirect evaluations for task-specific word embeddings – evaluating our causal models both *directly*, in terms of measuring their capacity to rank causally-related word pairs over word pairs of other relations, as well as *indirectly* in the downstream causal QA task. In both tasks, our analysis indicates that including causal models significantly improves performance. However, from the direct evaluation, it is difficult to estimate which models will perform best in real-world tasks. Our analysis re-enforces recent observations about the limitations of word similarity evaluations (Faruqui et al., 2016): we show that they have limited coverage and may align poorly with real-world tasks.

1.3.3 Contribution 3: Creating and ranking justifications for interpretable question answering

We propose a method to construct graph-structured answer justifications (text aggregation graphs) through aggregating relevant information from multiple sources.

Our graph structures model both intrasentence structures (i.e., syntax) and intersentence lexical overlap. Our empirical analysis demonstrates that modeling the contextual relevance of intersentence connections is crucial for good performance. This approach uses a latent-variable ranking perceptron algorithm that learns to jointly rank answers and justifications. In this extension of a traditional ranking perceptron, since we do not have labels for the *quality* of the justification, we model this as a latent variable to be learned. We evaluate our system on a large corpus of 1,000 elementary science exam questions from third to fifth grade, and demonstrate that our system significantly outperforms several strong learning-to-rank baselines at the task of choosing the correct answer. Further, we manually annotate answer justifications provided by the best baseline model and our intersentence aggregation method, and show that the intersentence aggregation method produces good justifications for 57% of questions answered correctly, significantly outperforming the best baseline method. Through an in-depth error analysis, we show that most of the issues encountered by the intersentence aggregation method center on solvable surface issues rather than complex inference issues. To our knowledge, this is the largest evaluation and most in-depth error analysis for explainable inference in the context of elementary science exams.

1.3.4 Contribution 4: Using neural networks to rank justifications for interpretable question answering

We propose an end-to-end neural method for learning to answer questions and select a high-quality justification for those answers. This approach re-ranks free-text answer justifications *without* the need for structured knowledge bases. With supervision only for the correct answers, we learn this re-ranking through a form of distant supervision – i.e., the answer ranking supervises the justification re-ranking. In this approach, we investigate two distinct categories of features in this "little data" domain: explicit features, and learned representations. We show that, with limited

training, explicit features perform far better despite their simplicity. We also demonstrate a large (+9%) improvement in generating high-quality justifications over a strong information retrieval baseline, while maintaining near state-of-the-art performance on the multiple-choice science-exam QA task, demonstrating the success of the end-to-end strategy.

1.4 Overview

The rest of this work is organized as follows. In Chapter 2 we outline the previous work relevant to our task. We then detail our word-association approaches to question answering that emphasize robustness: in Chapter 3 we present our alignment approach whereby we generate artifically aligned texts to serve as a proxy for aligned question-answer pairs; then in Chapter 4 we present our method for tailoring word-associations (here, in the form of word-embeddings) to a specific question-type. The next two chapters contain our methods which focus more on model-interpretability, providing a human-readable justification for each answer selected by the model. In Chapter 5 we detail an approach that uses aggregation to generate justifications for answers, extracts features based on a structured representation of the aggregated justiifcation, and then uses these features in a latent-variable reranking perceptron. The perceptron learns to simultaneously rank these justifications and answer candidates, thus providing the answer as well as the human-readable justification. In Chapter 6 we modify this approach to operate over much larger resources by removing the aggregation component as well as the structured representations, meanwhile extending the learning framework to a non-linear neural network. Finally, in Chapter 7 we discuss the work as a whole as well as future directions.

REFERENCES

Balduccini, M., C. Baral, and Y. Lierler (2008). Knowledge representation and question answering. *Foundations of Artificial Intelligence*, **3**, pp. 779–819.

Banarescu, L., C. Bonial, S. Cai, M. Georgescu, K. Griffitt, U. Hermjakob, K. Knight, P. Koehn, M. Palmer, and N. Schneider (2013). Abstract Meaning Representation for Sembanking. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pp. 178–186. Association for Computational Linguistics.

Baral, C. and S. Liang (2012). From Knowledge Represented in Frame-Based Languages to Declarative Representation and Reasoning via ASP. In *KR*.

Baral, C., S. Liang, and V. Nguyen (2011). Towards deep reasoning with respect to natural language text in scientific domains. In *DeepKR Workshop*. Citeseer.

Baral, C., N. H. Vo, and S. Liang (2012). Answering Why and How questions with respect to a frame-based knowledge base: a preliminary report. In *ICLP (Technical Communications)*, pp. 26–36. Citeseer.

Barzilay, R. and M. Lapata (2008). Modeling local coherence: An entity-based approach. *Computational Linguistics*, **34**(1), pp. 1–34.

Barzilay, R. and K. R. McKeown (2005). Sentence fusion for multidocument news summarization. *Computational Linguistics*, **31**(3), pp. 297–328.

Barzilay, R., K. R. McKeown, and M. Elhadad (1999). Information fusion in the context of multi-document summarization. In *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics*, pp. 550–557. Association for Computational Linguistics.

Berger, A., R. Caruana, D. Cohn, D. Freytag, and V. 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.

Björkelund, A. and J. Kuhn (2014). Learning structured perceptrons for coreference resolution with latent antecedents and non-local features. In *Proceedings of the Association for Computational Linguistics*.

Blair-Goldensohn, S., K. McKeown, and A. H. Schlaikjer (2003). A hybrid approach for answering definitional questions. *Technical Report CUCS-006-03, Columbia University*.

Bordes, A., S. Chopra, and J. Weston (2014). Question Answering with Subgraph Embeddings. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP) 2014*.

Bordes, A., N. Usunier, S. Chopra, and J. Weston (2015a). Large-scale simple question answering with memory networks. *arXiv preprint arXiv:1506.02075*.

Bordes, A., N. Usunier, S. Chopra, and J. Weston (2015b). Large-scale Simple Question Answering with Memory Networks. *CoRR*, **abs/1506.02075**.

Bordes, A., N. Usunier, A. García-Durán, J. Weston, and O. Yakhnenko (2013). Translating Embeddings for Modeling Multi-relational Data. In *NIPS*.

Brown, P. F., S. A. D. Pietra, V. J. D. Pietra, and R. L. Mercer (1993). The Mathematics of Statistical Machine Translation: Parameter Estimation. *Computational Linguistics*, **19**(2), pp. 263–311.

Brysbaert, M., A. Warriner, and V. Kuperman (2014). Concreteness ratings for 40 thousand generally known English word lemmas. *Behavior Research Methods*, **46**(3), pp. 904–911. doi:10.3758/s13428-013-0403-5.

Chen, D., J. Bolton, and C. D. Manning (2016a). A Thorough Examination of the CNN / Daily Mail Reading Comprehension Task. In *Proceedings of Association for Computational Linguistics (ACL)*.

Chen, D., J. Bolton, and C. D. Manning (2016b). A Thorough Examination of the CNN/Daily Mail Reading Comprehension Task. In *Association for Computational Linguistics (ACL)*.

Chen, D. and C. D. Manning (2014a). A Fast and Accurate Dependency Parser using Neural Networks. In *Proc. of the Conferenc on Empirical Methods for Natural Language Processing (EMNLP)*.

Chen, D. and C. D. Manning (2014b). A Fast and Accurate Dependency Parser using Neural Networks. In *Empirical MNLP*, pp. 740–750.

Chollet, F. (2015). Keras. `https://github.com/fchollet/keras`.

Chu-Carroll, J., K. Czuba, J. M. Prager, A. Ittycheriah, and S. Blair-Goldensohn (2004). IBM's PIQUANT II in TREC 2004. In *Text Retrieval Conference (TREC)*.

Clark, P. (2015). Elementary School Science and Math Tests as a Driver for AI: Take the Aristo Challenge! In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence, January 25-30, 2015, Austin, Texas, USA.*, pp. 4019–4021.

Clark, P., P. Harrison, and N. Balasubramanian (2013a). A study of the knowledge base requirements for passing an elementary science test. In *Proc. of the 2013 workshop on Automated Knowledge Base Construction (AKBC)*, pp. 37–42.

Clark, P., P. Harrison, and N. Balasubramanian (2013b). A study of the knowledge base requirements for passing an elementary science test. In *Proceedings of the 2013 Workshop on Automated Knowledge Base Construction*, AKBC'13, pp. 37–42.

Collins, M. (2002). Discriminative Training Methods for Hidden Markov Models: Theory and Experiments with Perceptron Algorithms. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, EMNLP '02, pp. 1–8. Association for Computational Linguistics, Stroudsburg, PA, USA. doi:10.3115/1118693.1118694.

Collobert, R., J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu, and P. Kuksa (2011). Natural language processing (almost) from scratch. *Journal of Machine Learning Research*, **12**(Aug), pp. 2493–2537.

Craven, M. W. and J. W. Shavlik (1996). Extracting tree-structured representations of trained networks. *Advances in neural information processing systems*, pp. 24–30.

Daumé III, H. (2007). Frustratingly Easy Domain Adaptation. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pp. 256–263. Association for Computational Linguistics, Prague, Czech Republic.

De Marneffe, M.-C. and C. D. Manning (2008). Stanford typed dependencies manual. Technical report, Technical report, Stanford University.

Dong, L., F. Wei, M. Zhou, and K. Xu (2015). Question answering over freebase with multi-column convolutional neural networks. In *Proceedings of Association for Computational Linguistics*, pp. 260–269.

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

Echihabi, A. and D. Marcu (2003b). A Noisy-Channel Approach to Question Answering. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 16–23. Sapporo, Japan.

Etzioni, O. (2011). Search needs a shake-up. *Nature*, **476**(7358), pp. 25–26.

Faruqui, M., Y. Tsvetkov, R. Rastogi, and C. Dyer (2016). Problems With Evaluation of Word Embeddings Using Word Similarity Tasks. *arXiv preprint arXiv:1605.02276*.

Feng, V. W. and G. Hirst (2012). Text-level Discourse Parsing with Rich Linguistic Features. In *Proceedings of the Association for Computational Linguistics*.

Fernandes, E. R., C. N. Dos Santos, and R. L. Milidiú (2012). Latent structure perceptron with feature induction for unrestricted coreference resolution. In *Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL) -Shared Task*, pp. 41–48. Association for Computational Linguistics.

Ferrucci, D., E. Brown, J. Chu-Carroll, J. Fan, D. Gondek, A. A. Kalyanpur, A. Lally, J. W. Murdock, E. Nyberg, J. Prager, et al. (2010). Building Watson: An overview of the DeepQA project. *AI magazine*, **31**(3), pp. 59–79.

Ferrucci, D. A. (2012). Introduction to "This is Watson". *IBM Journal of Research and Development*, **56**(3.4).

Finkel, J. R. and C. D. Manning (2010). Hierarchical joint learning: Improving joint parsing and named entity recognition with non-jointly labeled data. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pp. 720–728. Association for Computational Linguistics.

FitzGerald, N., O. Täckström, K. Ganchev, and D. Das (2015). Semantic Role Labeling with Neural Network Factors. In *Proc. of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 960–970.

Fried, D., P. Jansen, G. Hahn-Powell, M. Surdeanu, and P. Clark (2015). Higher-order Lexical Semantic Models for Non-factoid Answer Reranking. *Transactions of the Association for Computational Linguistics*, **3**, pp. 197–210.

Gondek, D., A. Lally, A. Kalyanpur, J. W. Murdock, P. A. Duboué, L. Zhang, Y. Pan, Z. Qiu, and C. Welty (2012). A framework for merging and ranking of answers in DeepQA. *IBM Journal of Research and Development*, **56**(3.4).

Graff, D., J. Kong, K. Chen, and K. Maeda (2003). English gigaword, LDC2003T05. *Linguistic Data Consortium, Philadelphia*.

Halliday, M. A. K. and R. Hasan (2014). *Cohesion in english*. Routledge.

Harabagiu, S., D. Moldovan, M. Pasca, R. Mihalcea, M. Surdeanu, R. Bunescu, R. Girju, V. Rus, and P. Morarescu (2000). Falcon: Boosting Knowledge for Answer Engines. In *Proceedings of the Text REtrieval Conference (TREC)*. Gaithersburg, MD, USA.

He, H. and J. Lin (2016). Pairwise word interaction modeling with deep neural networks for semantic similarity measurement. In *Proceedings of NAACL-HLT*, pp. 937–948.

He, X. and D. Golub (2016). Character-Level Question Answering with Attention. In *EMNLP*.

Hendrickx, I., S. N. Kim, Z. Kozareva, P. Nakov, D. Ó Séaghdha, S. Padó, M. Pennacchiotti, L. Romano, and S. Szpakowicz (2009). Semeval-2010 task 8: Multi-way classification of semantic relations between pairs of nominals. In *Proc. of the Workshop on Semantic Evaluations: Recent Achievements and Future Directions*, pp. 94–99.

Hermann, K. M., T. Kočiský, E. Grefenstette, L. Espeholt, W. Kay, M. Suleyman, and P. Blunsom (2015). Teaching Machines to Read and Comprehend. In *Advances in Neural Information Processing Systems (NIPS)*.

Hernault, H., H. Prendinger, D. duVerle, and M. Ishizuka (2010). HILDA: A Discourse Parser Using Support Vector Machine Classification. *Dialogue and Discourse*, **1**(3), pp. 1–33.

Hickl, A., J. Williams, J. Bensley, K. Roberts, B. Rink, and Y. Shi (2006). Recognizing textual entailment with LCCs GROUNDHOG system. In *Proceedings of the Second PASCAL Challenges Workshop*.

Hochreiter, S. and J. Schmidhuber (1997). Long short-term memory. *Neural computation*, **9**(8), pp. 1735–1780.

Hoffmann, R., C. Zhang, X. Ling, L. Zettlemoyer, and D. S. Weld (2011). Knowledge-based weak supervision for information extraction of overlapping relations. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, pp. 541–550. Association for Computational Linguistics.

Iyyer, M., J. Boyd-Graber, L. Claudino, R. Socher, and H. Daumé III (2014). A Neural Network for Factoid Question Answering over Paragraphs. In *Empirical Methods in Natural Language Processing*.

Iyyer, M., V. Manjunatha, J. Boyd-Graber, and H. Daumé III (2015a). Deep unordered composition rivals syntactic methods for text classification. In *Proceedings of the Association for Computational Linguistics*.

Iyyer, M., V. Manjunatha, J. Boyd-Graber, and H. Daumé III (2015b). Deep Unordered Composition Rivals Syntactic Methods for Text Classification. In *Association for Computational Linguistics*.

Jansen, P., N. Balasubramanian, M. Surdeanu, and P. Clark (2016). What's in an Explanation? Characterizing Knowledge and Inference Requirements for Elementary Science Exams. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pp. 2956–2965. The COLING 2016 Organizing Committee, Osaka, Japan.

Jansen, P., R. Sharp, M. Surdeanu, and P. Clark (2017). Framing QA as Building and Ranking Intersentence Answer Justifications. *Computational Linguistics*.

Jansen, P., M. Surdeanu, and P. 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)*.

Khashabi, D., T. Khot, A. Sabharwal, P. Clark, O. Etzioni, and D. Roth (2016). Question Answering via Integer Programming over Semi-Structured Knowledge. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence (IJCAI-16*, pp. 1145–1152.

Khoo, C. S., J. Kornfilt, R. N. Oddy, and S. H. Myaeng (1998). Automatic extraction of cause-effect information from newspaper text without knowledge-based inferencing. *Literary and Linguistic Computing*, **13**(4), pp. 177–186.

Khot, T., A. Sabharwal, and P. Clark (2017). Answering Complex Questions Using Open Information Extraction. In *Proceedings of Association for Computational Linguistics (ACL)*.

Kiela, D., F. Hill, and S. Clark (2015). Specializing Word Embeddings for Similarity or Relatedness. In *Proc. of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Kim, B., J. A. Shah, and F. Doshi-Velez (2015). Mind the Gap: A Generative Approach to Interpretable Feature Selection and Extraction. In *NIPS*.

Kingma, D. and J. Ba (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Lee, H., A. Chang, Y. Peirsman, N. Chambers, M. Surdeanu, and D. Jurafsky (2013). Deterministic coreference resolution based on entity-centric, precision-ranked rules. *Computational Linguistics*, **39**(4).

Lei, T., R. Barzilay, and T. S. Jaakkola (2016). Rationalizing Neural Predictions. In *EMNLP*.

Letham, B., C. Rudin, T. H. McCormick, D. Madigan, et al. (2015). Interpretable classifiers using rules and Bayesian analysis: Building a better stroke prediction model. *The Annals of Applied Statistics*, **9**(3), pp. 1350–1371.

Levy, O. and Y. Goldberg (2014). Dependency-Based Word Embeddings. In *Proc. of the 52nd Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 302–308.

Levy, O., S. Remus, C. Biemann, I. Dagan, and I. Ramat-Gan (2015). Do supervised distributional methods really learn lexical inference relations. In *Proc. of the Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*.

Lewis, M. and M. Steedman (2013). Combining Distributional and Logical Semantics. *Transactions of the Association for Computational Linguistics*, **1**, pp. 179–192.

Li, J., X. Chen, E. H. Hovy, and D. Jurafsky (2016). Visualizing and Understanding Neural Models in NLP. In *HLT-NAACL*.

Liang, P., A. Bouchard-Côté, D. Klein, and B. Taskar (2006). An end-to-end discriminative approach to machine translation. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, pp. 761–768. Association for Computational Linguistics.

Liang, P., M. I. Jordan, and D. Klein (2013). Learning dependency-based compositional semantics. *Computational Linguistics*, **39**(2), pp. 389–446.

MacCartney, B. (2009). *Natural language inference*. Ph.D. thesis, Citeseer.

Mann, W. C. and S. A. Thompson (1988). Rhetorical Structure Theory: Toward a Functional Theory of Text Organization. *Text*, **8**(3), pp. 243–281.

Manning, C. D., P. Raghavan, and H. Schütze (2008). *Introduction to Information Retrieval*. Cambridge University Press.

Manning, C. D., M. Surdeanu, J. Bauer, J. Finkel, S. J. Bethard, and D. McClosky (2014a). The Stanford CoreNLP Natural Language Processing Toolkit. In *Proc. of the 52nd Annual Meeting of the Association for Computational Linguistics (ACL)*.

Manning, C. D., M. Surdeanu, J. Bauer, J. Finkel, S. J. Bethard, and D. McClosky (2014b). The Stanford CoreNLP natural language processing toolkit. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pp. 55–60.

McSherry, F. and M. Najork (2008). Computing Information Retrieval Performance Measures Efficiently in the Presence of Tied Scores. In *30th European Conference on IR Research (ECIR)*. Springer-Verlag.

Mikolov, T., K. Chen, G. Corrado, and J. Dean (2013a). Efficient Estimation of Word Representations in Vector Space. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Mikolov, T., M. Karafiat, L. Burget, J. Cernocky, and S. Khudanpur (2010). Recurrent neural network based language model. In *Proceedings of the 11th Annual Conference of the International Speech Communication Association (INTERSPEECH 2010)*.

Mikolov, T., I. Sutskever, K. Chen, G. S. Corrado, and J. Dean (2013b). Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pp. 3111–3119.

Moldovan, D., C. Clark, S. Harabagiu, and D. Hodges (2007). Cogex: A semantically and contextually enriched logic prover for question answering. *Journal of Applied Logic*, **5**(1), pp. 49–69.

Moldovan, D., C. Clark, S. Harabagiu, and S. Maiorano (2003a). Cogex: A logic prover for question answering. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1*, pp. 87–93. Association for Computational Linguistics.

Moldovan, D., M. Paşca, S. Harabagiu, and M. Surdeanu (2003b). Performance Issues and Error Analysis in an Open-domain Question Answering System. *ACM Trans. Inf. Syst.*, **21**(2), pp. 133–154. ISSN 1046-8188. doi:10.1145/763693.763694.

Moldovan, D. I. and V. Rus (2001). Logic form transformation of wordnet and its applicability to question answering. In *Proceedings of the 39th Annual Meeting on Association for Computational Linguistics*, pp. 402–409. Association for Computational Linguistics.

Moschitti, A. (2004). A study on Convolution Kernels for Shallow Semantic Parsing. In *Proceedings of the 42th Annual Meeting of the Association for Computational Linguistics (ACL)*.

Moschitti, A. and S. Quarteroni (2011). Linguistic Kernels for Answer Re-ranking in Question Answering Systems. *Information and Processing Management: an International journal*.

Moschitti, A., S. Quarteroni, R. Basili, and S. Manandhar (2007). Exploiting Syntactic and Shallow Semantic Kernels for Question/Answer Classification. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 776–783. Prague, Czech Republic.

Murdock, J. W., J. Fan, A. Lally, H. Shima, and B. Boguraev (2012). Textual evidence gathering and analysis. *IBM Journal of Research and Development*, **56**(3.4).

Napoles, C., M. Gormley, and B. Van Durme (2012a). Annotated Gigaword. In *Proceedings of the Joint Workshop on Automatic Knowledge Base Construction and Web-scale Knowledge Extraction*, AKBC-WEKEX '12, pp. 95–100. Association for Computational Linguistics, Stroudsburg, PA, USA.

Napoles, C., M. Gormley, and B. Van Durme (2012b). Annotated gigaword. In *Proc. of the Joint Workshop on Automatic Knowledge Base Construction and Web-scale Knowledge Extraction*, pp. 95–100.

Och, F. J. and H. Ney (2003). A Systematic Comparison of Various Statistical Alignment Models. *Computational Linguistics*, **29**(1), pp. 19–51.

Oh, J.-H., K. Torisawa, C. Hashimoto, M. Sano, S. De Saeger, and K. Ohtake (2013). Why-Question Answering using Intra-and Inter-Sentential Causal Relations. In *The 51st Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 1733–1743.

Parikh, A. P., O. Täckström, D. Das, and J. Uszkoreit (2016). A Decomposable Attention Model for Natural Language Inference. In *EMNLP*.

Park, J. H. and W. B. Croft (2015). Using Key Concepts in a Translation Model for Retrieval. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '15, pp. 927–930. ACM, New York, NY, USA. ISBN 978-1-4503-3621-5. doi:10.1145/2766462.2767768.

Piaget, J. (1954). *The Construction of Reality in the Child*. Basic Books.

Pradhan, S. S., V. Krugler, S. Bethard, W. Ward, D. Jurafsky, J. H. Martin, S. Blair-Goldensohn, A. H. Schlaikjer, E. Filatova, P. A. Duboué, et al. (2002). Building a Foundation System for Producing Short Answers to Factual Questions. In *TREC*.

Reece, J., L. Urry, M. Cain, S. Wasserman, and P. Minorsky (2011). *Campbell Biology*. Pearson Benjamin Cummings. ISBN 9780321558237.

Ribeiro, M. T., S. Singh, and C. Guestrin (2016). "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In *HLT-NAACL Demos*.

Riedel, S., L. Yao, A. McCallum, and B. M. Marlin (2013). Relation extraction with matrix factorization and universal schemas. In *Proc. of Annual Meeting of the North American Chapter of the Association for Computational Linguistics (NAACL)*.

Riezler, S., A. Vasserman, I. Tsochantaridis, V. Mittal, and Y. 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)*, pp. 464–471.

Riloff, E. (1996). Automatically generating extraction patterns from untagged text. In *Proc. of the National Conference on Artificial Intelligence (AAAI)*, pp. 1044–1049.

Sachan, M., A. Dubey, and E. P. Xing (2016). Science Question Answering using Instructional Materials. In *The 54th Annual Meeting of the Association for Computational Linguistics*, p. 467.

Severyn, A. and A. Moschitti (2012). Structural relationships for large-scale learning of answer re-ranking. In *SIGIR*.

Severyn, A. and A. Moschitti (2013). Automatic Feature Engineering for Answer Selection and Extraction. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Severyn, A. and A. Moschitti (2015). Learning to Rank Short Text Pairs with Convolutional Deep Neural Networks. In *SIGIR*.

Severyn, A., M. Nicosia, and A. Moschitti (2013). Learning Adaptable Patterns for Passage Reranking. In *Proceedings of the Seventeenth Conference on Computational Natural Language Learning (CoNLL)*.

Sharma, A., N. H. Vo, S. Aditya, and C. Baral (2015). Towards Addressing the Winograd Schema Challenge-Building and Using a Semantic Parser and a Knowledge Hunting Module. In *Proceedings of the International Joint Conferences on Artificial Intelligence (IJCAI)*.

Sharp, R., P. Jansen, M. Surdeanu, and P. Clark (2015a). Spinning straw into gold. In *Proc. of the Conference of the North American Chapter of the Association for Computational Linguistics - Human Language Technologies*.

Sharp, R., P. Jansen, M. Surdeanu, and P. Clark (2015b). Spinning Straw into Gold: Using Free Text to Train Monolingual Alignment Models for Non-factoid Question Answering. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 231–237. Association for Computational Linguistics, Denver, Colorado.

Shen, L. and A. K. Joshi (2005). Ranking and Reranking with Perceptron. *Machine Learning. Special Issue on Learning in Speech and Language Technologies*, **60**(1), pp. 73–96.

Soricut, R. and E. Brill (2006). Automatic Question Answering Using the Web: Beyond the Factoid. *Journal of Information Retrieval - Special Issue on Web Information Retrieval*, **9**(2), pp. 191–206.

Soricut, R. 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*.

Suggu, S. P., K. N. Goutham, M. K. Chinnakotla, and M. Shrivastava (2016). Deep Feature Fusion Network for Answer Quality Prediction in Community Question Answering. *arXiv preprint arXiv:1606.07103*.

Sultan, M. A., S. Bethard, and T. Sumner (2014). Back to Basics for Monolingual Alignment: Exploiting Word Similarity and Contextual Evidence. *Transactions of the Association for Computational Linguistics*, **2**, pp. 219–230.

Sun, X., T. Matsuzaki, D. Okanohara, and J. Tsujii (2009). Latent Variable Perceptron Algorithm for Structured Classification. In *IJCAI*, volume 9, pp. 1236–1242.

Surdeanu, M., M. Ciaramita, and H. Zaragoza (2011). Learning to Rank Answers to Non-Factoid Questions from Web Collections. *Computational Linguistics*, **37**(2), pp. 351–383.

Surdeanu, M., T. Hicks, and M. 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*.

Tan, M., C. N. dos Santos, B. Xiang, and B. Zhou (2016). Improved Representation Learning for Question Answer Matching. In *ACL*.

Tari, L. and C. Baral (2006). Using AnsProlog with Link Grammar and WordNet for QA with deep reasoning. In *Information Technology, 2006. ICIT'06. 9th International Conference on*, pp. 125–128. IEEE.

Theano Development Team (2016). Theano: A Python framework for fast computation of mathematical expressions. *arXiv e-prints*, **abs/1605.02688**.

Tieleman, T. and G. Hinton (2012). Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude.

Tymoshenko, K. and A. Moschitti (2015). Assessing the Impact of Syntactic and Semantic Structures for Answer Passages Reranking. In *Proceedings of The 24th ACM International Conference on Information and Knowledge Management (CIKM)*.

Valenzuela-Escárcega, M. A., G. Hahn-Powell, and M. Surdeanu (2016). Odin's Runes: A Rule Language for Information Extraction. In *Proc. of the 10th International Conference on Language Resources and Evaluation (LREC)*.

Verberne, S., L. Boves, N. Oostdijk, P.-A. Coppen, et al. (2007). Discourse-based answering of why-questions. *Traitement Automatique des Langues, Discours et document: traitements automatiques*, **47**(2), pp. 21–41.

Voorhees, E. M. (2003). Evaluating Answers to Definition Questions. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology: Companion Volume of the Proceedings of HLT-NAACL 2003–short Papers - Volume 2*, NAACL-Short

'03, pp. 109–111. Association for Computational Linguistics, Stroudsburg, PA, USA. doi:10.3115/1073483.1073520.

Wang, D. and E. Nyberg (2015). A long short-term memory model for answer sentence selection in question answering. *ACL, July*.

Wang, M. and C. D. Manning (2010). Probabilistic Tree-Edit Models with Structured Latent Variables for Textual Entailment and Question Answering. In *COLING*.

Woodsend, K. and M. Lapata (2015). Distributed Representations for Unsupervised Semantic Role Labeling. In *Proc. of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Yang, M.-C., N. Duan, M. Zhou, and H.-C. Rim (2014). Joint Relational Embeddings for Knowledge-based Question Answering. In *Proc. of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 645–650.

Yao, X., B. Van Durme, C. Callison-Burch, and P. Clark (2013). Semi-Markov Phrase-based Monolingual Alignment. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Yih, W.-t., M.-W. Chang, X. He, and J. Gao (2015). Semantic parsing via staged query graph generation: Question answering with knowledge base. In *Proceedings of Association for Computational Linguistics (ACL)*.

Yih, W.-t., M.-W. Chang, C. Meek, and A. Pastusiak (2013). Question Answering Using Enhanced Lexical Semantic Models. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (ACL)*.

Zeiler, M. D. and R. Fergus (2014). Visualizing and Understanding Convolutional Networks. In *ECCV*.

Zettlemoyer, L. S. and M. Collins (2007). Online Learning of Relaxed CCG Grammars for Parsing to Logical Form. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pp. 678–687.