Composing Representations for Processes using Semantic Roles

Niranjan Balasubramanian

September 26, 2015

Contents

1 Introduction

Standardized test benchmarks such as grade-level science exams [3] motivate automatic knowledge extraction and reasoning [4, 5, 10]. Effective semantic representations of the relevant knowledge is critical for making significant advances in building such systems. In this proposal we focus on extracting knowledge about physical, chemical, biological processes and other natural phenomena. We propose a semantic-role based representation that effectively supports reasoning for answering grade-level questions on processes. Our proposal includes investigation of new methods for acquiring the typical semantic roles for processes, and methods for interpreting the roles mentioned in a sentence using the acquired knowledge.

Much of question answering and knowledge extraction research has focused on extracting and reasoning with simple facts. Open-domain QA and factoid question answering on the web has driven much of this research. In contrast grade-level science exams go beyond retrieval abilities and test understanding and use of knowledge to reason about specific scenarios [4]. The knowledge-requirements span a wide range including the ability to recognize a scenario and map elements of the scenario to known processes or phenomena. We refer to this as process recognition. Consider the following examples from an actual 4th grade science exam.

1. *When plants use stored sugar for energy, they go through a process called*
   (A) photosynthesis (B) transpiration (C) respiration (D) perspiration.
2. *A pot is heated on a stove. Which process causes the metal handle of the pot to also become hot?*
   (A) Conduction (B) Convection (C) Radiation (D) Combustion.

These questions test the ability to recognize a process based on the description of a scenario.

The first question tests knowledge about biological processes. *Photosynthesis* and *respiration* both involve sugar and energy. Photosynthesis converts light energy to sugar, whereas (cellular) respiration releases energy in the sugar by breaking it down. Not surprisingly these processes are described using similar words, which makes bag-of-words style reasoning unreliable. Understanding the different roles that energy and sugar play in these processes, and knowledge of the main actions involved is key to effective reasoning.

The second question tests knowledge about heat transfer mechanisms. The notion of *medium* is critical for answering this question. To perform effective reasoning a system must chain together few pieces of information: 1) The result of heating is that the pot becomes hot. 2) The pot handle is part of the pot and is thus in direct contact with the pot. 3) In a conduction process heat is transferred via direct contact. Most of these bits of information are naturally expressed via semantic roles of *heating* and *conduction*. e.g., *undergoer* = pot handle, *action* = heat transfer, *medium* = direct contact via part-of(pot handle, pot). Note that bits 1 and 2 are background knowledge and are not explicitly stated in the question. To apply these background knowledge we must be able to identify *what is being heated* and *what has become hot* in the scenario.

On the one hand, existing semantic resources such as FrameNet and PropBank provide different representations of general open-domain actions but do not cover knowledge about scientific processes. FrameNet, for instance, does not have entries for nearly half of the processes described in 4th grade science exams. The coverage is likely worse for higher grade levels. On the other hand, automatic knowledge bases built via relation extraction capabilities (e.g., Open IE) scale to arbitrary target concepts but only contain simple binary relations e.g., (Arg1, Rel, Arg2), that do not adequately capture the semantics.

In response, we propose to build a knowledge base about physical, chemical, and biological processes from their textual descriptions. The central idea is to automatically compose a semantic representation using a pre-determined vocabulary of semantic roles. Having identified the roles of interest, we will seek out sentences that express these roles and build extractors for these roles via bootstrapping. [1] We will conduct intrinsic and external evaluations. We will create a manual target representation for intrinsic evaluation, and use the 4th grade science questions as an external application. The proposed work has three major research components:

1.1 Representation – Discovering Roles

Our central premise is that the entities involved in a process and the roles they play provide a powerful representation for reasoning and QA. Similar representations have been shown to be useful for Open-domain factoid question answering [18, 16], and reading comprehension tasks on process descriptions [1].

As preliminary work, we first analyzed the knowledge requirements for a set of questions targeting around 150 processes. While a small collection of general purpose roles (e.g., Undergoer, Result, Enabler, Trigger) capture the key semantic elements for a majority of the processes [15], we also find that a set of domain specific roles (e.g., Direction, Medium, Physical_Property, Chemical_Property) are also critical. We propose to compose a representation that combines both general and domain specific roles.

Note that these roles are specified based on a canonical view of the process. One of the key challenges in question answering is in overcoming the variability of expression. Labeling with respect to a canonicalized view of the process is a deliberate attempt at addressing the variability problem.

Manually specifying roles for each process is not desirable from a scalability standpoint. We will investigate scalable methods for automatically identifying the appropriate roles for a process. Lexical and syntactic cues are indicative of certain roles in sentences. For example a *[process] happens by [event]* pattern suggests the presence of a manner role, and the presence of locative prepositions can indicate location or orientation roles. Note that the goal of this step is to simply assess whether a particular role applies to the process frame and not to accurately determine the role filler itself. [2] Furthermore, by aggregating these cues over large number of sentences, we can make reliable assessments about the importance of certain roles. [3]

1.2 Knowledge Extraction – Acquiring Role Fillers

Much of SRL work has focused on interpreting sentences i.e., identifying roles expressed in a sentence. Our goal is different. We seek to acquire knowledge about processes and represent them in terms of semantic roles. Prior work on SRL techniques have largely relied on supervised learning techniques for learning to identify roles []. However, obtaining training data is often difficult and laborious, especially for complex tasks such as SRL. Several prior approaches have used semi-supervised learning (SSL) approaches to address this issue.

We propose a different approach, where we explicitly find sentences that express the roles of interest and do a joint inference of role labels across all sentences. This allows us greater flexibility in choosing which sentences to use (we can maximize role coverage, diversity of instances etc.). Aligning the various spans within these sentences allows us to do collective labeling (similar in spirit to transduction learning).

[1] We call these extractors rather than semantic parsers, since the goal here is to build knowledge about these processes and not necessarily to build a parser that can reliably identify *all* semantic roles expressed in a sentence.

[2] Many cues are frame specific and using them in general for all frames can be problematic. However, the goal here is to be inclusive and determine which roles apply. It may be possible to over generate the frame with many plausible roles in this stage and prune it down in the next step if finding role fillers fails.

[3] Some roles might be left implicit as common sense knowledge. Those remain beyond the scope of this work.

1.3 Summary

We propose to develop methods for composing semantic role knowledge about processes. Successful completion of the work will result in the following contributions:

- Knowledge Representation – Method for automatically determining roles that apply for a target concept. This is distinct from existing work which either assumes that the roles are fully pre-specified or are induced fully automatically. Our proposal occupies a middle-ground that leverages cues

2 Background

2.1 Preliminary Work

2.2 Research Challenges and Proposed Work

This proposal aims to tackle the significant challenges in automatically constructing knowledge about processes. To address these challenges, we propose the following:

3 Representation

Our goal is to design a representation that meets the inferential needs in downstream applications. While we cannot anticipate all needs in advance, we turn to the the grade-level process questions for determining a role vocabulary. Our prior work showed that general purpose semantic roles, and a handful of domain-specific roles provide effective coverage for a majority of the recognition questions. We will extend these with roles from existing resources such as FrameNet and PropBank to construct a comprehensive role vocabulary.

Table 1 lists a standard list of semantic roles []. These roles are typically defined with respect to a predicate, most often a verb. However, there is substantial variability in discourse reflecting the diversity in situations involving the same processes. Consider the following sentences:

- When water evaporates it changes to water vapor.
- The process by which water from the oceans rises into the atmosphere as water vapor.

Both sentences describe instances of the process of evaporation. With respect to the corresponding predicates, Water vapor is the *Goal* in the first sentence, but is a *Patient* in the second. However, both are essentially the *Result* of the process evaporation.

This fragmentation is problematic for both interpreting situations and for drawing inferences about them. To avoid these problems we propose to define the roles with respect to a canonical description of the process. Further we aim to assemble sentences that express information in expected ways, thus allowing us flexibility in determining canonical lexical realizations.

Manual assignment of roles for every process is not scalable. Instead we propose to automatically determine the applicable subset of roles from this vocabulary.

4 ProcIteRoles: Architecture

ProcIterRoles allows for targeted iterative acquisition and refinement of process roles. The central idea is to collect high quality sentences and roles and iteratively expand the acquisition to include additional sentences

Table 1: Semantic Role Vocabulary

agent	Causal agent of an event.
instrument	The object used to accomplish the goal.
cause	Event or Agent that causes the event to happen.
experiencer	The entity undergoing the event.
recipient	The benefactor of the process outcome.
path	The path through which the process happens.
location	The place where the process happens.
measure	A quantity associated with the process.
theme	A process entity that doesn't undergo any change.

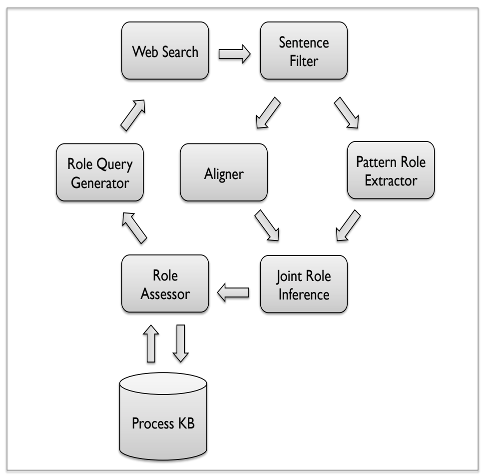

Figure 1: Process KB Acquisition: Proposed Architecture

and other roles.

Figure 1 shows the architecture of the acquisition process. Using simple query patterns we search for sentences that express roles with highly regular lexical cues. A sentence filter addresses sense issues and to remove malformed sentences. The filtered sentences are then processed through a pattern-based extractor that identifies and scores the candidate roles. In parallel, the sentences are also aligned to identify lexical units that should play similar semantic roles in the sentences. A joint inference module then provides a collective label assignment for all the sentences. The extracted roles are then assessed in conjunction with the existing roles. The assessor determines which roles should be added to the KB and also determines which roles need further additions. New query patterns are created for the roles that need addition and whole procedure is repeated again.

In the subsequent sections, we describe each component in greater detail.

5 Sentence Gathering

ProcIterRoles will leverage the vastness of the web to build a targeted collection of sentences that expresses roles of interest. There are two key challenges to address in gathering relevant sentences:

- **Relevance** – We want to find sentences that describe the target process of interest. The vastness of the web also means that there is information on nearly any possible interpretation of the words used to describe the process. For instance if we are interested in the process *crop fertilization*, we might also find information on *fertilization* in the reproductive sense, or in other metaphorical uses such as *cross fertilization of ideas*. Also, the target sense may not be a dominant sense on the web. Constructing effective keyword queries is therefore critical for finding relevant information.

- **Role Coverage** – We want to find sentences that cover all applicable roles that convey the desired information via simple expected constructions. Some roles are often expressed via highly regular lexico-syntactic constructions. On the other hand, roles such as *x* can be expressed in many different ways. Again with the vastness of the web, constructing effective role patterns is critical for finding useful sentences.

To address these challenges we propose two strategies. First, we target definitional sentences which convey the most salient information. Second, we espouse a feedback strategy that leverages high quality sentences gathered in the earlier phases to guide search in the later stages.

5.1 Querying

We propose to investigate simple but effective query pattern formulation methods. On a related task, our prior work had explored techniques that can find sentences that convey information about different aspects with respect to a topic (e.g., biographical aspects of a person). Our preliminary work shows that simple lexical templates e.g., "<process name> is the process by which" can yield high quality *definitional* sentences about processes. We use additional lexical templates for roles with regular lexicon-syntactic constructions e.g., "<process name> causes" is an effective pattern to find sentences that express the *result* roles of processes. The key challenge is to figure out querying patterns for roles with diverse lexical realizations. We propose adopting the standard bootstrapping approach used in relation extraction techniques to borrow functional patterns that introduce roles in other processes. Bootstrapping is known to introduce noise and topic drift issues. However, our approach doesn't entirely rely on the patterns alone. Rather we propose strong scoring and filtering mechanisms that can remove noise introduced via bootstrapping.

5.2 Scoring and Filtering

To account for the challenges in relevance, we seek to build a distributional context model that is seeded with some domain corpus. This model is then refined iteratively to allow for role coverage. Sentences from the web have high variance in quality and relevance.

6 Role Extraction

Many different approaches have been investigated for role labeling. The learning formulations studied range from pipelined classification approaches [9, 2], efficient structured and joint inference [12, 19], to end-to-end deep learning architectures [6]. Many different lexico-syntactic features, such as dependency paths and n-gram contexts, provide weak evidences for determining semantic roles [9]. Because these path-based and n-gram features are sparse, these supervised techniques require large amounts of training data. Semi-supervised and unsupervised approaches have been proposed as a means to address the training data problem [7, 11].

The focus of these approaches have been to build a SRL system that can identify the roles mentioned in a sentence. Our requirement is subtly different. We need to build a mechanism for acquiring the typical role fillers for a given process. First, we formulate a simpler local classification task that avoids the need for learning over role and predicate specific patterns. Starting with sentences that are likely to contain a specific role and a candidate text span from the sentence, we pose a classification task to determine if the candidate is indeed expressing role of interest. Then, we pose a joint inference task over multiple sentences, which allows us to use role decisions on similar text spans to influence each other.

6.1 Local Role Extraction

We set up a local (within sentence per role) extraction task. Relying on the patterns alone is problematic. Hand authored patterns, especially specifying the expected syntactic structure of the argument is quite limiting. Instead we generate many possible arguments that match a range of weakly indicative argument patterns and train a classifier.

The inputs are a sentence S, the role R for which it was retrieved, and the role pattern X_R that it matches, and a set of candidate spans C. The task is to predict if for each span if it is expressing the role of interest.

We adopt the standard SRL features such as clause, dependency path features, and n-gram context features [9, 12]. We explore two types of extensions that are specific to our setting:

- Different from a standard SRL setting, we seek identification of roles with respect to a canonical realization of the process. One can view this task as finding a mapping from predicate-specific semantic role to the process-specific role. To this end, we use an SRL system trained on PropBank data to identify predicate level semantic roles and use those as features to derive this mapping. Similarly, the frames that are evoked by the predicates in the sentence also provide important signals. For example, knowing that there is a conversion frame in a sentence increases the possibility of finding a result of a change of state of process like evaporation.

- Also we have strong expectations on *how* the argument is realized because the sentence is retrieved via a specific query pattern. This allows us to encode features that test if the expectations are met. [**Explain w/ an example**]

6.2 Joint Role Inference

Within sentence joint role inference has been shown to help SRL [17, 12]. Since our objective is to extract knowledge from multiple sentences, we propose to also exploit joint inference of roles across sentences. Local role extraction allows to reliably identify whether the specified role is expressed by the candidate text spans. However, this local classification is often inadequate because some cue patterns are ambiguous. For example, "evaporates into <x>" can match "steam" which is a *result* or "atmosphere" which is a *location*. Relying on the extraction pattern alone is problematic. Therefore, we propose to leverage role predictions on other similar text spans to improve inference.

Formalism

The key premise is that aligned text spans in similar sentences should be assigned same roles. This idea had been successfully used in semi-supervised and unsupervised settings to increase training data for SRL [7, 8, 14]. We adopt it for joint inference over test sentences, similar to the transductive learning approaches [**?**].

We extend an ILP-based formalism which has been shown to successfully model within sentence joint inference for SRL [17]. We add 1) a penalty term to the maximization objective that penalizes assignments that violate smoothness of labeling, 2) constraints that effectively fix labels from the local extractor for certain roles.

As noted earlier, the local extractor scores each text span $t_{i,k}$ from sentence S_k on how likely it is to belong to the role r_j. We use indicator variables to denote role assignment. z_{ijk} represents if the text span t_{ik} in sentence S_k is assigned the role r_j. Formally, the inference aims to find the best joint assignment to set of indicator variables F that maximizes the following objective function:

$$\arg\max_{\mathbf{z}} \sum_{i,j,k} z_{ijk} \cdot \rho(t_{ik}, r_j) \cdot \lambda_j - \beta \left\{ \sum_{i,k,l,m} \sigma(t_{ik}, t_{lm}) \left(\sum_{c \in |R|} |z_{ick} - z_{lcm}| \right) \right\}$$

$$\text{subject to}$$

$$\forall z_{ick} \in \mathbf{z}, \sum_{c=1}^{|R|} z_{ick} \leq 1 \text{[A span gets only one role.]}$$

$$\forall S_k \forall c \in |R|, \sum_{t_{ik} \in S_k} z_{ick} \leq 1 \text{[Roles are not repeated.]}$$

$$\cdots \text{[Other within sentence constraints.]}$$

Sentence and Text Span Alignment

The effectiveness of the joint inference relies on the ability to identify text spans in different sentences that should get the same role. Prior work explored a dependency graph-based approach to align predicate-argument structures in sentences. A linear combination of the overlap in lexical and syntactic structures of the candidate text spans is used to evaluate whether they should get the same roles [7, 8, 13]. This scoring function is used to transfer roles from a labeled sentence to an unlabeled sentence (semi-supervised setting) or to induce roles as clusters of arguments (unsupervised setting).

A key difference in our setting is that there are different sub-groups of sentences with different alignment characteristics. *Definition* sentences describe the process in terms of classes of entities and *instance* sentences which involve specific entities. Aligning definitional sentences is quite different from aligning definitions with instances or instances with other instances. Instances involve completely different entities which may not align via direct entailment but may align as substitutable siblings. Also, instances often tend to involve other non-essential information with respect to the process, whereas definitions are compact and tend to contain the most salient bits of information. To account for these differences we consider two extensions. Use different sets of weights and different similarity functions to combine the scores based on the types of sentences being aligned.

$$\sigma(t_{ik}, t_{lm}, u, v) = \alpha_{uv} \cdot lexsim_{uv}(t_{ik}, t_{lm}) + (1 - \alpha_{uv}) \cdot synsim_{uv}(t_{ik}, t_{lm})$$

[**Consider changing this to a trained classifier. Use sentence patterns rather than definition or instance sentence distinction.**]

7 Role Aggregation and Assessment

Inference yields a set of roles that can be reliably identified from the input set of sentences. The knowledge thus gathered is limited by the query patterns that we used to retrieve the sentences in the first place. To expand the knowledge further, we propose an iterative procedure that learns from the inferred roles.

First, given the current state of the knowledge base and newly inferred roles we devise a simple aggregation procedure that consolidates the roles – resolving any inconsistencies between the different iterations[4].

Next, we inspect the KB to identify roles that need to be filled, and pass them on to the sentence gathering components for expansion.

8 Evaluation

Our primary motivation is to compose knowledge that is necessary for grade-level science questions. Therefore we propose both an intrinsic evaluation that measures the accuracy of the extracted roles, and an external evaluation where we assess the utility of the roles in question answering.

8.1 KB Evaluation

Our goal is to acquire knowledge about processes along relevant role dimensions. [**To be filled...]**

8.2 Question Answering using Semantic Roles

Testbed Following Clark et al., 2014, we propose to build a large collection of multiple-choice questions at different grade levels. In prior work we leveraged existing collections to select questions. To effectively test the utility of our approach, we propose to create two types of questions in increasing levels of difficultly: 1) Questions that test ability to recognize instances of processes. 2) Questions that require ability to reason about the instance. We will limit the reasoning questions to those that are effectively achieved by following simple axioms defined over the semantic roles. [Example???]

Creating these questions is a difficult task for non-experts. We propose to work with the graduate students in the Department of Education at Stony Brook to collect questions and question templates. Once we have generate question templates that specify what kind of understanding is to be tested, we can scale out the question acquisition process via crowd sourcing.

Method

We use a simple approach to evaluate the utility of semantic roles for question answering. For the recognition questions we will follow our prior work, where we build a supervised role labeler and use it to parse the question to extract a role based representation. Then for each answer option we match the extracted roles against the roles available in the database.

For reasoning questions, we will utilize the above approach to first recognize the process and identify the roles mentioned in the question. We will utilize hand-written axioms defined over the roles. [Example.?]

9 Development Plan and Timeline

The project will proceed in three phases. In the first phase, we will design the representation and methods for role discovery and acquisition. In the second phase, we will start curating the generated knowledge and build a semantic role labeler for questions. In the third phase we will refine and make necessary adjustments

[4]It is possible to infer new roles with respect to the KB at each iteration, it can introduce many variables in inference and render it inefficient.

to the pipeline, finish the curation and release the resources. The research plan in calendar years is shown below:

Year 1:

1.
2.
3.

Year 2:

1.
2.
3.

10 Broader Impact

11 Curriculum Development Activities

I plan to teach a course centered around the core concepts of knowledge representation, and scalable extraction techniques for knowledge. This course is relevant for both Masters and PhD students. Most NLP-based technology companies and technology companies with a large web presence have a need for extracting and organizing knowledge from their user engagement data. This course will provide a basic overview of a distributed information extraction pipeline, persistence, and building applications that rely on the extracted data.

12 Prior NSF Support

Dr. Niranjan Balasubramanian has broad expertise in information extraction, esp. in large scale knowledge generation and its application to complex NLP tasks but has not received prior support from NSF.

References

[1] Jonathan Berant, Vivek Srikumar, Pei-Chun Chen, Abby Vander Linden, Brittany Harding, Brad Huang, Peter Clark, and Christopher D. Manning. Modeling biological processes for reading comprehension. In *Proceedings of EMNLP*, 2014.

[2] Anders Björkelund, Love Hafdell, and Pierre Nugues. Multilingual semantic role labeling. In *Proceedings of the Thirteenth Conference on Computational Natural Language Learning: Shared Task*, pages 43–48. Association for Computational Linguistics, 2009.

[3] Peter Clark. Elementary school science and math tests as a driver for ai: Take the aristo challenge. *to appear*, 2015.

[4] Peter Clark, Niranjan Balasubramanian, Sumithra Bhakthavatsalam, Kevin Humphreys, Jesse Kinkead, Ashish Sabharwal, and Oyvind Tafjord. Automatic construction of inference-supporting knowledge bases. In *AKBC2014*, Montreal, Canada, December 2014.

[5] Peter Clark, Phil Harrison, and Niranjan Balasubramanian. A study of the AKBC/requirements for passing an elementary science test. In *Proc. of the AKBC-WEKEX workshop at CIKM*, 2013.

[6] William R Foland Jr and James H Martin. Dependency-based semantic role labeling using convolutional neural networks. In *Proceedings of the Fourth Joint Conference on Lexical and Computational Semantics*, pages 279–288, 2015.

[7] Hagen Fürstenau and Mirella Lapata. Graph alignment for semi-supervised semantic role labeling. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 11–20, Singapore, 2009.

[8] Hagen Fürstenau and Mirella Lapata. Semi-supervised semantic role labeling via structural alignment. *Computational Linguistics*, 38(1):135–171, 2012.

[9] Daniel Gildea and Daniel Jurafsky. Automatic labeling of semantic roles. *Computational linguistics*, 28(3):245–288, 2002.

[10] Tushar Khot, Niranjan Balasubramanian, Eric Gribkoff, Ashish Sabharwal, Peter Clark, and Oren Etzioni. Exploring markov logic networks for question answering. In *Empirical Methods in Natural Language Processing*, 2015.

[11] Ivan Titov Alexandre Klementiev. Semi-supervised semantic role labeling: Approaching from an unsupervised perspective. In *Proceedings of the COLING Conference.*, 2012.

[12] Peter Koomen, Vasin Punyakanok, Dan Roth, and Wen-tau Yih. Generalized inference with multiple semantic role labeling systems. In *Proceedings of the Ninth Conference on Computational Natural Language Learning*, pages 181–184. Association for Computational Linguistics, 2005.

[13] Joel Lang and Mirella Lapata. Unsupervised induction of semantic roles. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 939–947, Los Angeles, California, June 2010. Association for Computational Linguistics.

[14] Joel Lang and Mirella Lapata. Unsupervised semantic role induction with graph partitioning. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1320–1331, Edinburgh, Scotland, UK., July 2011. Association for Computational Linguistics.

[15] Samuel Louvan, Chetan Naik, Veronica Lynn, Ankit Arun, Niranjan Balasubramanian, and Peter Clark. Semantic role labeling for process recognition questions. In *K-CAP Scientific Knowledge Workshop*, 2015.

[16] Luiz Augusto Pizzato and Diego Mollá. Indexing on semantic roles for question answering. In *Coling 2008: Proceedings of the 2nd workshop on Information Retrieval for Question Answering*, pages 74–81. Association for Computational Linguistics, 2008.

[17] Vasin Punyakanok, Dan Roth, Wen-tau Yih, and Dav Zimak. Semantic role labeling via integer linear programming inference. In *Proceedings of the 20th International Conference on Computational Linguistics*, COLING '04, Stroudsburg, PA, USA, 2004. Association for Computational Linguistics.

[18] Dan Shen and Mirella Lapata. Using semantic roles to improve question answering. In *EMNLP-CoNLL*, pages 12–21, 2007.

[19] Oscar Täckström, Kuzman Ganchev, and Dipanjan Das. Efficient inference and structured learning for semantic role labeling. *Transactions of the Association for Computational Linguistics*, 3:29–41, 2015.