Specifying and Annotating Reduced Argument Span Via QA-SRL

Gabriel Stanovsky Meni Adler Ido Dagan
Computer Science Department, Bar-Ilan University
{gabriel.satanovsky,meni.adler}@gmail.com
dagan@cs.biu.ac.il

Abstract

Prominent semantic annotations take an inclusive approach to argument span annotation, marking arguments as full constituency subtrees. Some works, however, showed that identifying a reduced argument span can be beneficial for various semantic tasks. While certain practical methods do extract reduced argument spans, such as in Open-IE , these solutions are often ad-hoc and system-dependent, with no commonly accepted standards. In this paper we propose a generic argument reduction criterion, along with an annotation procedure, and show that it can be consistently and intuitively annotated using the recent QA-SRL paradigm.

1 Introduction

Representations of predicate-argument structure need to determine the span of predicates and their corresponding arguments. Surprisingly, there are no accepted NLP-standards which specify what the "right" span of an argument should be.

Semantic representations typically take an inclusive (or maximal) approach: PropBank annotation (Palmer et al., 2005), for example, marks arguments as full constituency subtrees. From an application perspective, this maximal approach ensures that all arguments are indeed embedded within the annotated span, yet it is often not trivial how to accurately recover them.

In contrast to this maximal-span approach, Open-IE systems (Etzioni et al., 2008; Fader et al., 2011) put emphasis on extracting readable standalone propositions, typically producing shorter arguments (see examples in Section 2.1). Several recent works have exploited this property, using Open-IE extractions as an intermediate representation within a larger framework.

Angeli et al. (2015) built an Open-IE system which focuses on shorter argument spans. They hypothesize that "shorter arguments [are] more likely to be useful for downstream applications", and demonstrate this by using their system to extract facts about predefined entities in a state-of-the-art Knowledge Base Population system.

Further, Stanovsky et al. (2015) compared the performance of several off-the-shelf parsers in different semantic tasks. Most relevant to this work is the comparison between Open-IE and SRL. Specifically, they suggest that SRL's longer arguments introduce noise which hurts performance for downstream tasks. This is sustained empirically by showing that extractions from Open-IE4[1] significantly outperform ClearNLP's SRL (Choi, 2012) in textual similarity, analogies, and reading comprehension tasks.[2]

While Open-IE extractors do provide a reduction of argument span, they lack consistency and principled rigor – there is no clear definition for the desired argument span, which is defined de-facto by the different implementations. This lack of a common system-independent definition, let alone an annotation methodology, hinders the creation of gold standard argument-span annotation.

In this work we propose a concrete argument span reduction criterion and an accompanying annotation procedure, based on the recent QA-SRL paradigm (He et al., 2015). We show that this criterion can be consistently annotated with high agreement, and that it is intuitive enough to be obtained through crowd-sourcing.

As future work, we intend to apply the reduction criterion to other types of predicates (e.g., nomi-

[1] http://knowitall.github.io/openie

[2] Open IE-4 is based on ClearNLPs SRL, allowing for a direct comparison.

Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, pages 474–478,
Berlin, Germany, August 7-12, 2016.

nal and adjectival predication). Subsequently, we would like to create a comprehensive annotated resource, as a benchmark for the detection of reduced argument spans.

2 Background

2.1 Argument Span

As discussed in the Introduction, PropBank takes an inclusive approach to annotating arguments, by marking them as full constituency subtrees. For example, given the sentence "*Obama, the newly elected president, flew to Russia*", PropBank will mark "Obama, the newly elected president" as ARG0 of the predicate *flew*.

However, in certain applications, such as question answering or abstractive summarization, a reduced argument is preferred (i.e., "Obama"). Notably, different implementations of Open-IE provide an applicable generic way to reduce argument span. Since there are no common guidelines for this task, each Open-IE extractor produces different argument spans. We cover briefly some of the main differences in a few prominent Open-IE systems.

ReVerb (Fader et al., 2011) uses part-of-speech-based regular expressions to decide whether a word should be included within an argument span. For example, they move certain light verb compliments and prepositions from the argument to the predicate slot (e.g., "***gave a talk at***"). OLLIE (Mausam et al., 2012) learns lexical-syntactic patterns and splits extractions across certain prepositions. For example, given "*I flew from Paris to Berlin*", OLLIE yields (I; **flew**; from Paris) and (I; **flew**; to Berlin). More recently, (Angeli et al., 2015) used natural logic to remove non-integral parts of arguments (e.g., removing the underlined non-restrictive prepositional phrase in "*Heinz Fischer of Austria*").

2.2 QA-SRL

SRL is typically perceived as answering **argument role questions**, such as *who, what, to whom, when*, or *where*, regarding a target predicate. For instance, PropBank's ARG0 for the predicate *say* answers the question "*who said something?*".

QA-SRL (He et al., 2015) follows this perspective, and suggests that answering explicit role questions is an intuitive means to solicit predicate-argument structures from non-expert annotators. Annotators are presented with a sentence in which a target predicate[3] was marked, and are requested to annotate argument role questions (from a restricted grammar) and corresponding answers.

For example, given the previous sentence and the target predicate *flew*, an annotator can intuitively provide the following QA pairs: (1) *Who flew somewhere?* **Obama**, and (2) *Where did someone fly?* **Russia**.

The annotation guidelines further solicit multiple shorter answers, each typically embedded in the span of a maximal PropBank-style argument, while providing a different answer to the (same) argument role question.

In Section 4 we make use of QA-SRL's framework in order to produce annotations by our reduction argument criterion, which is defined in the next section.

3 Argument Reduction

In this section, we propose annotation criteria and process for obtaining minimal argument spans. Given an original, non-reduced argument, we aim to reduce it to a set of (one or more) smaller arguments, which jointly specify the same answer to the argument's role question.

Formally, given a non-reduced argument $a = \{w_1, ..., w_n\}$, along with its role question $Q(a)$ with respect to predicate p in sentence s, we seek to find a set of minimally-scoped arguments, $M(a)$, such that:

(1) Each $m \in M(a)$ is a proper subset of a.

(2) Each $m \in M(a)$ provides a different, independently interpreted answer to $Q(a)$.

(3) $M(a)$ is *equivalent* to a, in the sense that when taken jointly, $M(a)$ specifies the same answers as a does for $Q(a)$.

(4) Each $m \in M(a)$ is *minimal*, meaning it cannot be further reduced without violating the equivalence criterion (3).

Note that this definition relies on human judgments, which are used to decide whether two arguments provide the same or different answers.

Generally speaking, a non-minimal argument a can be reduced in one of two ways:

(a) *Removal* of tokens from a, forming a smaller argument.

[3] Currently these consist of automatically annotated verbs.

(b) *Splitting* a, yielding multiple arguments.

In our context, we would like to apply these two operations as long as they maintain the equivalence criterion (3). We empirically observe that the first case (removal) corresponds to the omission of non-restrictive modifiers, that is, modifiers for which the content of the modifier presents a separate, parenthetical unit of information about the NP (Huddleston et al., 2002). For example, revisiting the sentence: "*Obama, the newly elected president, flew to Russia.*", the non-reduced argument "Obama, the newly elected president" can be reduced to the minimal argument "Obama", as both specify the same answer to the role question "*who flew to Russia?*".

In contrast, a restrictive modifier is an integral part of the meaning of the containing NP, and hence should not be removed, as in "*She wore the necklace* ***that her mother gave her***".

The second reduction operation (splitting) corresponds to decoupling distributive coordinations, that is, cases in which a predicate applies separately to all of the elements in the coordination. For example, in: "*Obama and Clinton were born in America.*", the non-reduced PropBank-style argument "Obama and Clinton" can be reduced to two arguments {"Obama", "Clinton"}. Each of these arguments independently answers the role question "*Who was born in America?*", while jointly they correspond to the longer, non-reduced argument.

Note that splitting a shorter distributive argument does not necessarily produce disjoint arguments. For example, consider: "*The tall boys and girls were born in America.*", in which "The tall boys and girls" would reduce to two overlapping arguments: {"The tall boys", "The tall girls"}.

In contrast, non-distributive conjuncts cannot be split. These are cases in which the predicate applies to the conjuncts taken together, while applying it separately to each element changes the interpretation of the clause. Consider for example the reciprocal structure of: "*Obama and Putin met in Moscow*", in which we cannot split the argument "Obama and Putin" since the predicate *met* implies that Obama and Putin met with each other, which will be lost if we split the argument to two independent answers.

Based on these two operations, a set of minimal arguments, $M(a)$, can be obtained from a in a top-down manner: first apply *removal*, if possible; then *splitting*, if possible.[4] Next, apply recursively to each of the smaller arguments, stopping when none of the two reduction operations can be applied.

This annotation process might yield different sets of minimal arguments by different annotators, depending on their decisions regarding the reduction steps. As we show empirically in the next section, high agreement levels can be obtained, supporting the validity of our proposed criterion.

4 Annotation Experiment

In this section we describe the compilation and analysis of a small-scale expert annotation corpus. Creating such corpus serves 3 goals: (1) It allows us to test the applicability of the argument reducing procedure, (2) By comparing it with Propbank we can examine how often, and in which cases, we reduce arguments (Section 4.1), and (3) We can assess the plausibility of crowd-sourcing argument span annotation (Sections 4.2 and 4.3).

In order to achieve these goals, we sample 100 predicates of the Propbank corpus, which covered 260 arguments. To allow comparisons, we sample predicates which were annotated by QA-SRL and whose arguments were aligned by (He et al., 2015) with a matching Propbank argument.[5]

Two expert annotators used the QA-SRL's interface to re-answer the original QA-SRL annotated questions with minimally-scoped arguments, according to the procedure described in Section 3. Prior to annotating the expert dataset, the annotators discussed the process and resolved conflicts on a separate development set of 20 predicates.

Annotator agreement From an argument perspective, the annotators fully agreed on the span of 94.6% of the arguments.

Looking into the word token level, we found that for a given PropBank argument $a = (w_1, ..., w_n)$, the respective reduced arguments always constitute a subset of a. This allows us to look at the annotation process as a list of n mapping decisions – for each w_i, an annotator decides whether he (1) Maps it to one or more of the argu-

[4] This order is arbitrary, chosen solely to provide a deterministic process. Alternating the steps would yield an identical set.

[5] An annotated answer is judged to match the PropBank argument if either (1) the gold argument head is within the annotated answer span, or (2) the gold argument head is a preposition and at least one of its children is within the answer span.

ments of $M(a)$, or (2) Deletes it. The complete annotation required each annotator to make 985 such mappings decisions. Word level agreement between the annotators was calculated as the percent of the decisions on which they agreed, and found to be 97.1%.

Overall, the annotators achieved a high level of agreement, suggesting that the reduction criterion can be consistently applied by trained annotators. An analysis of the few disagreements revealed that the deviations between the annotators stem from semantic ambiguities, where two legitimate readings of the sentence led to different span annotations.[6]

Finally, we compose the expert annotation dataset from 247 arguments on which both annotators fully agreed.

4.1 Comparison with Propbank

Comparing our annotation with PropBank showed that we reduced roughly 24% of the arguments: 19% of the arguments were reduced by omitting non-restrictive modifications and 5% of the arguments were split across distributive co-ordinations (see discussion on both types of reductions in Section 3).

The average reduced argument shrunk by roughly 58%. In general, these numbers suggest that our annotation scheme targets commonly recurring phenomena, and significantly deviates from PropBank's annotation of arguments.

4.2 Crowdsourcing

We created an Amazon Mechanical Turk[7] project to investigate the possible scalability of our annotation using non-trained annotators.

Similarly to the setting used by the expert annotators, turkers were presented with a sentence, followed by a list of questions regarding a target predicate. The sentences, predicates and questions were taken from the expert corpus, which aligns between QA-SRL and Propbank.[8]

The guidelines for annotators refined those of He et al. (2015), soliciting answers which follow our formal criterion. In cases of multiple answers referring to the same entity, annotators are asked to provide the most specific answer, otherwise (if the answers refer to different entities), the annotators are asked to list all of the answers. Furthermore, the annotators are requested to provide the shortest answer they can, while preserving its correctness.

We chose annotations which were agreed upon by at least two annotators. In cases where the three annotators gave different answers (26% of the time), we used a fourth annotator to arbitrate, and calculated agreement using the same metrics discussed above. Cases where annotators disagreed were mostly semantically ambigouos. For example, given the sentence "*Our pilot simply laughed , fired up the burner and with another blast of flame lifted us , oh , a good 12 - inches above the water level .*" and the question "*how much did someone lift someone?*", one annotator replied **12 - inches** while another replied ***a good* 12 - inches**.

We found that the crowdsourcing annotations to be of high quality, reaching 89.1% argument agreement and 93.5% word agreement with our expert annotation. These results suggest that the annotation of argument span is efficiently and accurately attainable using crowd-sourcing techniques, with only subtle refinements over the original QA-SRL guidelines.

Annotation	Argument	Word
Expert - IAA	94.6%	97.1%
QA-SRL - Expert	80%	88.5%
Our Crowdsourcing - Expert	**89.1%**	**93.5%**

Table 1: Agreement levels between the different annotations: (1) IAA - Inter-Annotator agreement between the expert annotators (2) Agreement of QA-SRL corpus with our expert annotation and (3) Our Crowdsourcing - agreement of the Amazon Mechanical Turk annotations with our expert annotation. See Section 4.

4.3 Comparison with QA-SRL

Finally, we want to compare our crowdsourcing annotation versus that of QA-SRL, with respect to argument span. Using the previously mentioned agreement metric, we find that QA-SRL agrees with our expert dataset on 80% of the arguments and 88.5% of the word-level decisions. Although it is outperformed by our crowdsourcing annota-

[6] For example, in "*The American Stock Exchange said a seat was sold for $ 160,000 , down $ 5,000 from the previous sale last Friday .*", one annotator did not reduce ARG1, while the second annotator chose to restrict the span of the argument to "*a seat was sold for $ 160,00*", interpreting the remaining part of the clause as an addition by the author.

[7] https://www.mturk.com

[8] To be clear, the annotators saw only the raw text and questions from QA-SRL and were not exposed to the PropBank annotations.

tion project, QA-SRL still manages to capture significant amounts of the minimally-reduced arguments. This is interesting, as the QA-SRL guidelines did not address this issue specifically, but instead solicited annotators to provide "as many answers as possible". This suggests that the question answering format intuitively prompts human annotators to reduce the span of their answers.

To conclude this section, the entire comparison measurements are summarized in Table 1.

5 Conclusion and Future Work

In this work we proposed a concrete criterion for specifying minimally-scoped arguments. While this issue was applicably addressed by previous work, it was not consistently defined or annotated. Following this definition, we created an expert annotation dataset over texts from PropBank, using the QA-SRL paradigm. This annotation achieved high levels of inter-annotator agreement, and was shown to be intuitive enough so that it can be scaled to crowdsourcing annotation. As future work, we plan to extend this annotation project to larger volumes of text, and to additional types of (non-verbal) predications, which will allow to develop learning-based methods that identify minimally-reduced argument span.

Acknowledgments

We would like to thank Luheng He and Luke Zettlemoyer for the fruitful discussions, and the anonymous reviewers for their helpful comments.

This work was supported in part by grants from the MAGNET program of the Israeli Office of the Chief Scientist (OCS), the Israel Science Foundation grant 880/12, and the German Research Foundation through the German-Israeli Project Cooperation (DIP, grant DA 1600/1-1).

References

Gabor Angeli, Melvin Johnson Premkumar, and Christopher D. Manning. 2015. Leveraging linguistic structure for open domain information extraction. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics (ACL 2015)*.

Jinho D. Choi. 2012. *Optimization of Natural Language Processing Components for Robustness and Scalability*. Ph.D. thesis, Boulder, CO, USA. AAI3549172.

Oren Etzioni, Michele Banko, Stephen Soderland, and Daniel S Weld. 2008. Open information extraction from the Web. *Communications of the ACM*, 51(12):68–74.

Anthony Fader, Stephen Soderland, and Oren Etzioni. 2011. Identifying relations for open information extraction. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1535–1545. Association for Computational Linguistics.

Luheng He, Mike Lewis, and Luke Zettlemoyer. 2015. Question-answer driven semantic role labeling: Using natural language to annotate natural language. In *the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Rodney Huddleston, Geoffrey K Pullum, et al. 2002. The cambridge grammar of english. *Language. Cambridge: Cambridge University Press*.

Mausam, Michael Schmitz, Stephen Soderland, Robert Bart, and Oren Etzioni. 2012. Open language learning for information extraction. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 523–534, Jeju Island, Korea, July. Association for Computational Linguistics.

Martha Palmer, Daniel Gildea, and Paul Kingsbury. 2005. The proposition bank: An annotated corpus of semantic roles. *Computational linguistics*, 31(1):71–106.

Gabriel Stanovsky, Ido Dagan, and Mausam. 2015. Open IE as an intermediate structure for semantic tasks. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics (ACL 2015)*.