Automating FrameNet Annotation

Miriam R. L. Petruck
International Computer Science Institute
miriamp@icsi.berkeley.edu

Swabha Swayamdipta
Allen Institute for Artificial Intelligence
swabhas@allenai.org

July 2019

1 FrameNet

Background

FrameNet (Ruppenhofer et al., 2016) is a research and resource development project in corpus-based computational lexicography grounded in the principles of **frame semantics** (Fillmore, 1985). One of the goals of this effort is documenting the **valences**, i.e., the syntactic and semantic combinatorial possibilities of each item analyzed. These valence descriptions provide critical information on the mapping between form and meaning; such mapping is what Natural Language Processing (NLP) and Natural Language Understanding (NLU) require.

At the heart of the work is the **semantic frame**, a script-like knowledge structure that facilitates inferencing within and across events, situations, states-of-affairs, relations, and objects. FrameNet defines a semantic frame in terms of its **frame elements** (FEs), or participants in the scene that the frame captures; a **lexical unit** (LU) is a pairing of a lemma and a frame, thus characterizing that LU in terms of the frame that it evokes. Valence descriptions derive from the annotation of frame elements, i.e. **semantic roles**, on example sentences that illustrate the linguistic manifestation of the participants in a scene with respect to the target of analysis. Frame Elements are in fact triples of information; they also provide information about the grammatical function (GF) and the phrase type (PT) of the annotated FE. In FrameNet, automatic processes recommend GFs and PTs, which annotators correct, when necessary.

Figure 1 displays annotation for a sentence with one target LU, its corresponding frame, FEs, GFs, and PTs. The verb `bought` is the target of analysis; the annotations instantiate the COMMERCE_BUY

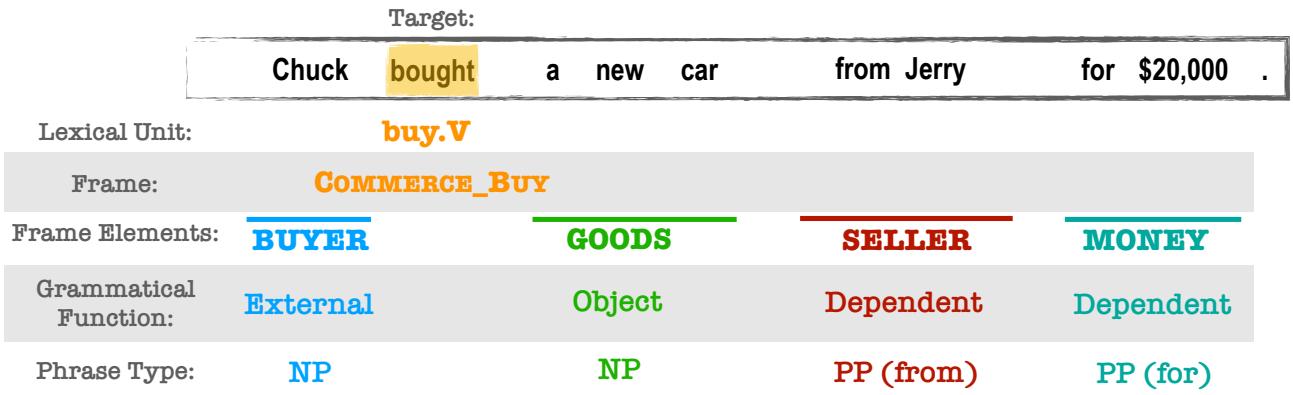

Figure 1: FrameNet Annotation

frame, whose FEs, BUYER, SELLER, MONEY, and GOODS, annotators *manually* label on the appropriate constituents in the sentence.

Current Status

FrameNet has defined 1,224 frames, hosting 13,640 lexical units, for which it has provided over 202,000 annotation (sets). In addition, the FrameNet hierarchy links related frames (to each other) using 1,876 frame-to-frame relations via more than 10,725 frame element relations.[1]

Of the 13,640 lexical units in the FrameNet database, only 62% have associated annotations; thus, 38% of the lexical units in the database do not have associated annotations. The goal of this work is to investigate and develop ways to automate semantic role labeling (FE annotation) for the 38%, with an eye toward increasing coverage in FrameNet by using automatic processes to produce annotated data for use in NLP and NLU.

2 Semantic Role Labeling

Background

Semantic Role Labeling (SRL) is the task of automatically identifying semantic frame elements for a given target lexical unit - frame pair in any given sentence. Gildea and Jurafsky (2002)'s pioneering work, based on early FrameNet data (Johnson et al., 2002), paved the way for the development of SRL systems, some of which show promise for a high level of performance (e.g., Swayamdipta et al., 2018; Roth and Lapata, 2016, 2015; Das et al., 2014). Several systems have made use of semantic frames and frame elements for improved performance in various NLP application (e.g., Agarwal et al., 2014; Liu et al., 2016).

Recently Developed Systems

This investigation will employ three recently developed systems.

- `SEMAFOR`[2] (Das et al., 2014) uses a pipeline of discrete, manually designed feature-based classifiers for target identification, frame identification, and argument identification. Integer linear programming based constraints are applied to satisfy the core properties of frame-semantic analyses, such as non-repetition of core frame elements, etc.

- PathLSTM[3] (Roth and Lapata, 2016) uses neural features for embedding path relationships between frames and frame elements, in a pipeline similar to that of `SEMAFOR`.

- `open-SESAME`[4] (Swayamdipta et al., 2018) uses an unconstrained, neural approach in a similar pipeline. Continuous representations, as well as a sophisticated global model based on semi-Markov conditional random fields improve the identification of arguments.

The use of multiple systems offers two benefits: first, we plan to use the proposed annotations for which consensus among the systems exists; secondly, we might be able to compare these different systems on the basis of their linguistic judgment or prediction power of the system, and generalizability of the system to new data.

[1] `https://framenet.icsi.berkeley.edu/fndrupal/current_status`
[2] `https://github.com/Noahs-ARK/semafor`
[3] `https://github.com/microth/PathLSTM`
[4] `https://github.com/swabhs/open-sesame`

3 The Challenge of Manual Annotation

The NLP community recognizes the value that FrameNet holds (e.g. Smith, 2017). For instance, research on linguistic structure prediction exploits FrameNet data for the development of tools (e.g., frame-semantic parsing) essential for a variety of NLP applications, such as machine translation, question-answering, information extraction, and text summarization.

Manual annotation is time-consuming and expensive; nevertheless, the ever-increasing need for more data persists. Advances in technology, including better performing SRL systems (e.g. Swayamdipta et al., 2018; Roth and Lapata, 2016), and FrameNet 's current status demand investigating the viability of incorporating automatic processes into FrameNet 's development process, rather than only relying on manual annotation. The popularity of SRL at major computational linguistics conferences testifies to the community's interest in the task. What remains is the annotation of illustrative example sentences, the data that computational linguists and NLP developers need. To date, all of FrameNet's FE data has been annotated manually.

Recent Efforts in FrameNet

FrameNet has provided undergraduate students with research experience in the context of the project. These students received mentoring and participated in the many aspects of FrameNet's work, be that any high-level discussion about the analysis of data, or some practical matters such as updating the bibliography or improving the usability of the project's public website.[5]

Since 2016, FrameNet has hosted three different students, each interested in SRL (also as a computational task, independent of its value for FrameNet), and each working on developing software for FrameNet to run `SEMAFOR` or (more recently) `open-SESAME` locally. Since undergraduate students tend only to work with FrameNet for a semester, or in rare cases, a year, FrameNet has yet to achieve even this initial goal, although FrameNet is inching closer.

We believe that the time is right for FrameNet to address incorporating SRL into the FrameNet process in a systematic way, and without the limitation of unavoidable discontinuous work.

The Challenge: Can FrameNet develop reliable ways of producing more annotated data in a more efficient and cost-effective manner than before?

4 Rising to the Challenge

This investigation will concentrate on the 38% of lexical units in the FrameNet database without associated annotation. Since FrameNet has assigned these lexical units to existing frames manually, with highly skilled FrameNet lexicographers performing the work, the word-sense disambiguation part of FrameNet analysis is done. Because these lexical units have been characterized in terms of frames with other lexical units that do have associated annotations, the possibility of taking advantage of these annotations exists.

This idea corresponds to a zero-shot learning (Socher et al., 2013) for argument identification, since targets and frames have previously been identified. As frame elements or roles are fixed for a given frame according to its FrameNet definition, prior knowledge in the form of annotations for other lexical units for the same frame could be leveraged for this zero-shot learning paradigm. Moreover, knowledge about frame elements can be obtained from FE-FE mappings of *related* frames, based on the FrameNet hierarchy; this information would provide indirect supervision for the prediction of semantic roles for the desired lexical units. Since GFs and PTs are obtained automatically for the original annotations, these can also be important features for the prediction of frame elements for new lexical units. Existing

[5]Depending on the the nature of the project, students can receive course credit for their work.

FrameNet-internal resources form an interesting application of low resource machine learning (Gormley et al., 2014), since on average *annotated* lexical units contains approximately 20 manual annotations.

5 Closing the Scientific Loop

The goals of this project go well beyond automatic collection of annotations for FrameNet. Again, using FrameNet data, Gildea and Jurafsky (2002) developed the first-ever SRL system. That is, FrameNet's manually annotated data motivated (in part) Gildea and Jurafsky (2002)'s development of SRL; that work also initiated SRL as a now well-recognized task in the field. In turn, advances in SRL systems over recent years and the need for more semantically rich annotated data motivate this investigation. Enriching FrameNet and collecting annotations will inspire some current SRL systems to be more powerful; and in turn the sufficiently mature systems could be tested for their viability in FrameNet's ongoing development.

Thus, in addition to the likelihood of FrameNet benefiting from such systems, SRL systems would also benefit from a larger FrameNet—in effect closing the scientific loop. Furthermore, this effort would open up the path to an even more ambitious zero-shot learning problem for frames without any annotations in the FrameNet database; this work also has implications for low-resource NLP. Importantly, assuming a positive outcome to this study, incorporating SRL into FrameNet's process would "return the results" of the project's earliest work back into the project itself.

References

Apoorv Agarwal, Sriramkumar Balasubramanian, Anup Kotalwar, Jiehan Zheng, and Owen Rambow. 2014. Frame semantic tree kernels for social network extraction from text. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 211–219.

Dipanjan Das, Desai Chen, André F. T. Martins, Nathan Schneider, and Noah A. Smith. 2014. Frame-semantic parsing. *Computational Linguistics*, 40(1):9–56.

Charles J. Fillmore. 1985. Frames and the semantics of understanding. *Quaderni di Semantica*, 6(2):222–254.

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

Matthew R Gormley, Margaret Mitchell, Benjamin Van Durme, and Mark Dredze. 2014. Low-resource semantic role labeling. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1177–1187.

Christopher R Johnson, Charles J Fillmore, Esther J Wood, Josef Ruppenhofer, Margaret Urban, Miriam RL Petruck, and Collin F Baker. 2002. The FrameNet project: Tools for lexicon building.

Shulin Liu, Yubo Chen, Shizhu He, Kang Liu, and Jun Zhao. 2016. Leveraging FrameNet to improve automatic event detection. In *Proc. of ACL*, pages 2134–2143.

Michael Roth and Mirella Lapata. 2015. Context-aware frame-semantic role labeling. *Transactions of the Association for Computational Linguistics*, 3:449–460.

Michael Roth and Mirella Lapata. 2016. Neural semantic role labeling with dependency path embeddings. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1192–1202.

Josef Ruppenhofer, Michael Ellsworth, Miriam R. L. Petruck, Christopher R. Johnson, Collin F. Baker, and Jan Scheffczyk. 2016. *FrameNet II: Extended Theory and Practice*. ICSI: Berkeley.

Noah Smith. 2017. Squashing computational linguistics. Association for Computational Linguistics.

Richard Socher, Milind Ganjoo, Christopher D Manning, and Andrew Ng. 2013. Zero-shot learning through cross-modal transfer. In *Advances in neural information processing systems*, pages 935–943.

Swabha Swayamdipta, Sam Thomson, Kenton Lee, Luke Zettlemoyer, Chris Dyer, and Noah A. Smith. 2018. Syntactic scaffolds for semantic structures. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3772–3782, Brussels, Belgium. Association for Computational Linguistics.