PropBank Comes of Age—Larger, Smarter, and more Diverse

Sameer Pradhan[1,2], Julia Bonn[3], Skatje Myers[3], Kathryn Conger[3].
Tim O'Gorman[4], James Gung[5], Martha Palmer[3]
[1]University of Pennsylvania, Philadelphia, PA, 19104, [2]cemantix.org
[3]University of Colorado, Boulder, CO 80303,
[4]Thorn, MI (thorn.org), [5]Amazon AI, New York City, NY
pradhan@cemantix.org, martha.palmer@colorado.edu

Abstract

This paper describes the evolution of the PropBank approach to semantic role labeling over the last two decades. During this time the PropBank frame files have been expanded to include non-verbal predicates such as adjectives, prepositions and multi-word expressions. The number of domains, genres and languages that have been PropBanked has also expanded greatly, creating an opportunity for much more challenging and robust testing of the generalization capabilities of PropBank semantic role labeling systems. We also describe the substantial effort that has gone into ensuring the consistency and reliability of the various annotated datasets and resources, to better support the training and evaluation of such systems.

1 Introduction

Twenty years ago traditional statistical machine learning techniques were holding sway and successful stochastic syntactic parsing was on the rise. The availability of accurate syntactic parses opened the door to richer, deeper representations. The second Human Language Technology conference included a presentation on *Adding Predicate Argument Structure to the Penn Treebank* and the Proposition Bank (PropBank) was born (Kingsbury and Palmer, 2002). Over the next few years, with the able guidance of a steering committee consisting of Ralph Weischedel, Mitch Marcus, Doug Appelt, Mark Villain and Ralph Grishman, the annotation guidelines and the annotation continued to grow, with the end result of over 110,000 predicate argument structures pointing directly to syntactic nodes in the phrase structure syntax trees of the roughly 50,000 sentences of the Penn Treebank. The annotation of these structures was guided by a set of approximately 3300 Frame Files that provided a verb specific set of semantic roles as the arguments for each verb. The substantial size of the data set and the consistency of the annotation gave rise to a flurry of popular semantic role labeling systems and semantic role labeling shared tasks (Carreras and Màrquez, 2005; Surdeanu et al., 2008) that continue to this day. The Penn Treebank is entirely composed of Wall Street Journal articles, and annotation of additional data taken from the more diverse English genres of the Brown corpus allowed for out of domain testing, with predictable dismal results. Since that time, DARPA and NSF have funded substantial additional PropBank annotation, focusing on additional domains and genres for English, as well as additional languages such as Chinese, Arabic, Hindi and Urdu. The deep learning revolution has not abated the interest in semantic role labeling performance, and the incorporation of PropBank Frame files into the Abstract Meaning Representation (AMR) Editor (Banarescu et al., 2013), to guide the labeling of the AMR nested predicate argument structures, ensures its longevity. This paper details the new genres, domains and datasets that are now available, as well as the expansion of the original PropBank verb Frame Files to adjectival and nominal forms. Today PropBank has a prominent web presence[1] and plans to evolve and cater to the growing, global, distributed, diverse community by means of a GitHub organization[2]. Github supports the infrastructure for streamlining contributions and resolving issues that are bound to arise in the future. Multiple versions of stable annotations are made available to the community for promoting open, reproducible research[3]. Different versions of the frame lexicon can be viewed and searched online in a human friendly format[4].

We start by reviewing the framework and assumptions for the original PropBank and detail

[1] http://propbank.org

[2] http://github.com/propbank

[3] Many diverse sources and subcorpora are covered by the sum total of all annotations. Access to various data slices is governed by the data and privacy restrictions on the underlying source. A bulk of the data is accessible free of charge for research use upon completion of relevant data use paperwork. The details can be found on the main website.

[4] http://propbank.org/v3.4.0/frames

Proceedings of the 11th Joint Conference on Lexical and Computational Semantics, pages 278 - 288

the changes that have been made as it matured in Section 2. Section 3 describes the additional new domains and genres that are covered in subsequent annotation efforts. In Section 4 we provide novel baseline results on these new corpora to stimulate additional research in the robustness and portability of semantic role labeling. Finally we summarize our contributions in Section 5.

2 The Proposition Bank, Then and Now

PropBank (Kingsbury and Palmer, 2002; Palmer et al., 2005a) is a paradigm for the development of corpora annotated with predicate argument structures. In its original form, these predicate arguments structures were applied to the syntactic scaffolding provided by the Penn Treebank. While creating a global inventory of semantic roles was traditionally viewed as too difficult, PropBank sidestepped the issue by using an "individual thematic roles" approach (Dowty, 1991) in which roles are custom-defined within each (coarse-grained) sense of a predicate. The decision as to what constitutes a semantic role and the use of Penn Treebank as the syntactic scaffolding for the annotation contributed to high inter-annotator agreement, which led to higher performing machine learning models and fueled interest in the task. PropBank has been instrumental in creating a subfield of NLP called Semantic Role Labeling (SRL). The following three subsections describe the evolution of PropBank in terms of the kind of predicates that were annotated, the changes seen in the data structures as paradigm matured, and the genre of data annotated over the past two decades.

2.1 Frames—Predicate Rolesets and Arguments

The core of the PropBank paradigm consists of an annotation schema and a lexical inventory collectively referred to as the **Frames**. Frames are a set of files that house "rolesets", which are predicate argument structures associated with coarse-grained senses for eventualities. Within a roleset, roles that are considered semantically and/or syntactically core are bundled together as predicate-specific numbered arguments[5]. In annotation, all rolesets across all predicates share a larger pool of "adjunct" arguments such as ARGM-LOC for location, ARGM-TMP for temporal, ARGM-GOL for goals and beneficiaries, etc. These three-letter ARGM tags cover generalized thematic role information that is more specific than argument numbers but more categorical than custom role definitions, and so the pool of ARGMs has become the basis for a set of *function tags* that are now applied to roles. The list of function tags includes PAG (proto-agent) and PPT (proto-patient), taken from Dowty (1991); the list of function tags continues to grow along with PropBank's expansion into more domain-specific corpora. Each role in every roleset comes with an argument number, a custom definition, and a function tag as described in Figure 1.

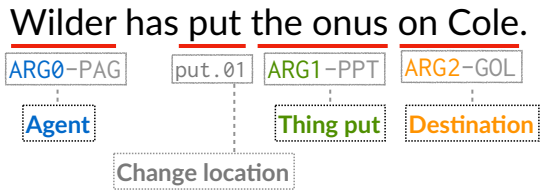

Figure 1: In this example, the verb predicate `put` invokes the *change of location* roleset `put.01` in which the proto-agent (PAG) is the numbered argument ARG0 getting assigned a value ARG0-PAG; *Thing put*, is the proto-patient (PPT), getting value ARG1-PPT; and *the destination* being a goal (GOL) getting the value ARG2-GOL.

The Frames are not in themselves organized according to any kind of semantic hierarchy; rolesets are grouped inside frame files according to polysemy and etymological closeness and nothing more (e.g. the `leave` frame file includes rolesets for multiple `leave` and `left` predicates). However, each roleset potentially includes links with other lexical resources such as VerbNet (Schuler, 2005), FrameNet (Baker et al., 1998), etc., as well as to word senses in WordNet (Fellbaum, 2010) and therefore to those in OntoNotes (Weischedel et al., 2011; Pradhan et al., 2013). This collectively forms a rich, interconnected, high coverage, semantic network.

Over time, the significance of the lexicon of predicate frames has risen to the level where it is *not just an artifact of PropBank*, but has become a resource in its own right, forming the backbone of various meaning representations such as AMR, Uniform Meaning Representation (UMR) (Gysel et al., 2021), etc.

[5]We use the term "argument" when referring to the general notion of arguments of a predicate; and the terms "role" and "rolesets" when we are referring to the vocabulary of roles assigned to each argument of a predicate in the lexicon of (mostly lemma specific) frames.

2.2 Coverage—Genres and Languages

The original PropBank comprised a single news genre, as represented by the WSJ. Over time more genres and languages were PropBanked. At first a small subset of the Brown corpus was annotated to test the generalizability of machine learning models. Subsequently, as part of the OntoNotes project, it covered more genres and was adapted to two other languages—Chinese (Palmer et al., 2005b) and Arabic. The OntoNotes genres include broadcast news, broadcast conversations, web text (blogs, newsgroups), telephone conversations (Godfrey et al., 1992; Taylor, 1996), and a pivot corpus of New and Old Testament text.

The methodology has been adapted to Korean (Palmer et al., 2006), Hindi/Urdu (Bhatt et al., 2009), Finnish (Haverinen et al., 2013), Turkish (Sahin, 2016), Persian (Mirzaei and Moloodi, 2016), Russian (Moeller et al., 2020), and Brazilian Portuguese (Duran and Aluísio, 2011).

PropBank was further extended to additional languages by the Universal PropBanks (Akbik et al., 2015; Jindal et al., 2022). Some of these were automatically generated by projecting English SRL annotation onto parallel text in seven languages and further refining them through filtering and bootstrapping.

2.3 Evolution of the Data Structure

The first version of the PropBank was annotated on top of constituent trees of the Penn Treebank. As a result, with a few exceptions, the PropBank semantic role labels represent nodes in a constituent parse tree. As PropBank grew, it uncovered areas in the Treebank guidelines that conflicted with the PropBank semantic interpretation choices. This led to an effort to synchronize the two resources, creating an improved version of each (Babko-Malaya et al., 2006). Initial machine learning approaches converted the annotation into a series of text spans (Carreras and Màrquez, 2005) and relied heavily on a syntactic parser for good performance (Pradhan et al., 2005). The period starting around 2007 saw a significant rise in the use of dependency representation of parses. International evaluations of dependency parse based semantic role labeling were originally organized by automatically mapping the constituent tree semantic roles to dependency trees (Surdeanu et al., 2008). In the last decade, thanks in part to a combination of the advent of deep learning and the maturity of the guidelines and existing models, PropBank annotations have been freed from the syntactic scaffolding provided by the Treebank. The more recent PropBank annotations are performed on flat text[6]. The core lexicon for PropBank which are the frame files follow an XML specification which has evolved through several iterations over the years. All annotations have been updated to match the latest version of the specification.

2.3.1 Why XML and not JSON?

Contrary to popular notion, JSON is NOT universally better than XML[7]. In fact, as this three part series of articles[8,9,10] highlights, as of now, XML schema[11] is still the most versatile form of defining and validating declarative data specifications and constraints when compared to its JSON counterpart—JSON Schema. We are currently in the process of moving away from a somewhat restrictive DTD specification to a full-fledged XML schema. We could consider migrating to the JSD(x) which uses a JSON schema definition language (JSD) modeled closely with XML Schema language and guarantees a one-to-one mapping between the two[12]

2.4 Frames—Updated Specification

2.4.1 Synchronizing with AMR

The first release of PropBank only covered verbal predicates. Nominal forms in the Penn Treebank were handled by the NomBank project at NYU (Meyers et al., 2004). During the OntoNotes project, the PropBank Frame Files were expanded to include eventive nominals such as NomBank nominalizations, which had already been based on the original verb frame files, as well as light verb constructions (Hwang et al., 2010). By 2012, in support of the Abstract Meaning Representation (AMR) project, PropBank introduced other non-verbal predicates including additional noun

[6] More details regarding the evolution of annotation file formats can be found in the documentation available on the PropBank website.

[7] This subsection added to address a reviewer concern.

[8] https://www.toptal.com/web/json-vs-xml-part-1

[9] https://www.toptal.com/web/json-vs-xml-part-2

[10] https://www.toptal.com/web/json-vs-xml-part-3

[11] Two expressive XML schema languages are in widespread use—XML Schema (with a capital S) and RELAX NG.

[12] The combination JSD and JSDx—shortened as JSD(x)—is a self-describing schema where the language JSD(x) is expressed in JSD(x) itself and allows declarative specification of structural and functional constraints equivalent to XML schema. Moving from XML schema to JSONx should be quite straight-forward when the supporting infrastructure reaches a reasonable level of maturity.

forms, adjectives, and certain multi-word expressions. AMR's aim was to abstract away from syntactic specificity and annotate semantic argument structures for eventualities regardless of their part of speech. Initially, new predicate types were added as distinct rolesets–for example, `fear-n.01` (noun) and `afraid-j.01` (adjective) were modeled after `fear-v.01` (verb), sharing its semantics and argument structure but operating independently (Bonial et al., 2014, 2012, 2017). While the new additions more than tripled the range of what was annotatable, they also introduced a certain amount of redundancy into the lexical inventory, and so the entire lexicon was put through an extensive overhaul to unify etymologically-related rolesets, increasing the similarity to FrameNet frames. The 2017 post-unification release introduced a new roleset structure in which multiple predicates (aliases) could be included in a single roleset (e.g. `fear.01`, with aliases `fear-v`, `fear-n`, and `afraid-j`) (O'Gorman et al., 2018). It also introduced new varieties of complex multi-word predicates including multi-word expressions (MWEs)—fully noncompositional idioms like `jump_the_shark` as well as semi-decompositional expressions like `have_in_mind`—and predicating prepositional phrases like `in_love`.

In the five years since the post-unification release, PropBank's lexical inventory has been recruited for an increasingly broad range of domain-specific annotation projects across PropBank and AMR. With each of these projects comes a unique set of annotation needs that have broadened the scope of the lexical inventory. For example, the Spatial AMR annotation project expands the PropBank lexicon and AMR annotation schema to allow for grounded annotation of multimodal spatial corpora (Bonn et al., 2020; Narayan-Chen et al., 2019). The particular needs of the project meant expanding the rolesets to allow non-eventuality predicate types, like prepositional relations and their etymologically-related adverbial counterparts (e.g. spatial direction terms like `back`, `left`. While not eventualities, such expressions still benefit from the sense disambiguation and essential role clustering that come with roleset treatment. Because grounded annotation of directed spatial expressions requires tracking the linguistic frame of reference of each instance, these spatial rolesets are also the first in the PropBank lexicon to introduce numbered arguments for roles that are essential yet almost never explicitly realized.

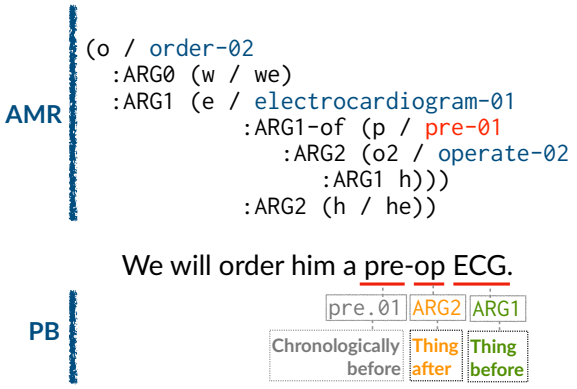

Figure 2: This example shows how the frame lexicon is shared between two representations—Abstract Meaning Representation (AMR) and PropBank (PB) in the clinical domain. The predicate `pre` invokes the *Chronologically before* roleset `pre.01`, where, the *Thing before* is assigned the role ARG1 and the *Thing after* is assigned the role ARG2. Note that the AMR shows the arguments of three additional predicate rolesets: i) `order-02`; ii) `electrocardiogram-01`; and iii) `operate-02`, which correspond to what PropBank would annotate as predicates for the tokens having the surface forms "order", "ECG," and "op" respectively.

The THYME project is another domain-specific AMR annotation project that has required significant, specialized expansion of the lexical inventory. The THYME colon cancer corpus consists of cancer-related clinical-narrative documents that have been annotated in such a way as to provide temporal-relation extraction of clinical events (Albright et al., 2013; Styler et al., 2014; Wright-Bettner et al., 2019). The corpus contains highly specialized medical terminology rarely seen in the general domain: surgical procedures, anatomical parts, diseases, disorders, symptomatology, etc. One of the great challenges of this project has been to determine which of these types to treat as unnamed (decomposable) entities, which to treat as named entities, and which to treat with rolesets. The emphasis on temporal relations in THYME reveals that concepts that would not formerly have been considered eventive enough to qualify for roleset treatment do in fact function as eventulaties in medical corpora, with complex argument structures that need to be tracked even when implicit. THYME is also responsible for adding PropBank's first affix rolesets for temporally-indicative prefixes like `pre-` and `post-`, which, for temporal relation purposes, need to be annotated separately from their stem events. An example[13] of this is shown in Figure 2.

[13]There is a slight notational difference between AMR us-

With the needs of domain-specific annotation projects pulling further and further from the center of the original framing guidelines, PropBank has been overdue for an update that emphasizes flexibility for domain-of-use. While the inventory continues to exist as a single cohesive whole, we have added structures into the files that allow users to extract rolesets that are associated with domain-specific projects. For general domain rolesets, we have also made it easier to identify ways in which the roleset may be applied differently from one project to another, including usage tags as well as expanded examples that showcase differences in annotation strategies for different projects. Rolesets may now include aliases from any of the following parts of speech as required by a project: verb (`v`), noun (`n`), adjective (`j`), LVC[14] (`l`), MWE (`m`), preposition (`p`), adverb (`r`), and affix (`f`). There may now also be aliases (`argaliases`) associated with numbered arguments (e.g., ARG0 of `teach.01` may have an `argalias` of `teacher`). The next section describes how these changes are manifest in the Frames' `xml` files..

2.4.2 Enriched Contents

This latest PropBank 3.4.0 release[15] uses an enriched `xml` specification which provides some additional features and allows for better validation and disambiguation.

Lexlink Tags We aim to provide mappings between PropBank and other lexical resources within the frame files themselves, when available. The `<lexlinks>` tag provides correspondences between a given roleset and equivalents in VerbNet, FrameNet, or OntoNotes senses. The `<rolelinks>` tag additionally provides mappings between specific roles and these external resources.

For example, `sing.01` includes `<lexlink>` tags linking the roleset to both `manner_speaking-37.3` and `sound_emission-43.2` in VerbNet 3.4. The `<rolelink>` tags on the ARG1 specify that it is the equivalent of *topic* and *theme* for those two VerbNet classes, respectively.

Usage Tags The updated version also now includes `<usage>` tags to specify whether they were included during the development of a particular version of a resource. Many rolesets were constructed only for use with AMRs (`*-91`), and some only for a particular project, such as Spatial AMRs. Table 1 lists the various values and the corpora they correspond with. Within the repository, we provide a utility script that can reduce the XML files down to only the rolesets included in a specified resource/version.

Example Tags The `<example>` tags within the frame files have had a major overhaul. In order to accommodate AMRs, these tags now use `<propbank>` to contain the PropBank annotations for an example sentence and `<amr>` to contain the AMR graph. Additionally, the `<amr>` tag may specify the version of AMR, as AMR projects may annotate the same text in different ways.

Example sentence text now comes with the expectation of being tokenized. The `<arg>` and `<rel>` tags previously only required specification of the text that should be annotated, but this allowed for ambiguous interpretations. If an argument was a word that showed up multiple times in the sentence, there was no way to clarify which instance was the correct argument. The improved format requires the specification of start/end indices for annotated spans. Not only does this prevent ambiguity, it also allows for machine reading/validation of the examples created by human annotators, such as ensuring that arguments do not overlap.

Additionally, `<arg>` tags within the examples now use a single `type` attribute to specify the role, such as ARG0 or ARGM-MNR. This primarily serves to improve readability of the XML compared to the previous `f` and `n` attributes used to specify the same information.

One of the most significant changes to the examples is transforming them to a syntax-agnostic format. Previously, examples in the frame files used a variety of syntactic notation to aid annotators using the constituent-parse-based Jubilee tool with the expectation of a regimen of post-processing. Arguments were frequently noted to be a syntactic trace, such as `*trace*`. We have eliminated these by either resolving them to their true text span or removing the argument entirely if it is only implied but not present in the text. Converting the examples to this more generalized format greatly improves readability and adaptability for new projects or annotation schemes that don't depend on phrase structure parses.

MWE tags Multi-word expressions that receive mappings between literal and figurative meanings

ing a hyphen instead of a period separating the lemma and the roleset.

[14] Light Verb Construction

[15] Henceforth we are going to follow the SemVer (semantic versioning) scheme: https://semver.org/

Resource	Version	Description
PropBank	3.4.0	Latest release
PropBank	Flickr 1.0	Flickr captions dataset
PropBank	3.1	Unification release (ON, BOLT, LORELEI)
		English Web Treebank (EWT)
PropBank	2.1.5	OntoNotes v5.0 (ON)
PropBank	1.0	Proposition Bank I
AMR	2019	General-purpose AMR rolesets
AMR	THYME 1.0	THYME colon cancer corpus
AMR	Spatial 1.0	Minecraft Dialogue Corpus

Table 1: Resource/version combinations present in the <usage> tags.

have changed format as well. In the previous release, an <mwe> tag inside the <aliases> tag housed elements describing the <tokens> involved in the expression, and a <mappings> tag that was sister to <aliases> housed the source to target semantic mappings. The new version renames <mwe> as <mwp-descriptions>, places the <tokens> inside a new element called <syntaxdesc>, and pulls the <mappings> in so that all MWE-related information is contained in one place in the file.

2.4.3 Quality control

The format overhaul required significant examination of the current data. Through a combination of conservative automated processes and extensive manual correction, the new release offers consistency that previously was unavailable and impractical. Subsequent releases will benefit from both these corrections and a format more compatible with future machine validation. We are in the process of updating the way the proposition layer is serialized. The original version was a file with a prop extension which contained one predicate argument structure per line, and where the predicate and arguments were identified using pointers to node(s) in the Treebank parse of the sentence containing the predicate. The new serialization will no longer be so tightly coupled with the nodes in the parse tree[16].

The new release updates examples to current PropBank guidelines. Outdated SLC and RCL roles have been updated to use the current R- argument convention. In the sentence "The acre of ground *that* adjoins our property.", the relativizer *that* used to be annotated with ARGM-SLC, which was linked to the span *the acre of ground* (tagged as ARG1 of predicate *adjoins*). This was an artifact of the strong alignment of PropBank role (spans) to nodes in the syntactic parse tree and required an additional processing step. The annotation for the relativizer is now tagged as R-ARG1. Examples that used ARGA were too sparse and infrequent and have been updated and that role has been eliminated.

As part of validating the frame examples, we've corrected numerous cases caused by human error, such as examples missing a <rel> tag, the specified argument text not corresponding with the sentence text, or multiples of the same numbered argument.

Within the repository, we provide a script to perform a validation check on a directory of frame files. This includes not only checking the XML format according to the DTD, but other common sense checks, such as that example arguments' indices correspond correctly with the example text, that arguments don't overlap, and ensuring the same numbered argument isn't present multiple times.

2.4.4 Available Tools

We previously named two scripts to help users work with the frame files: one that provides validation checks and another that can pare the XML files down to only rolesets included in a particular resource. These scripts are available on the git repository.

Additionally, we provide a script that can be used to generate a user-friendly website based on this new format of XML files. The website provides searchability based on roleset ID or alias, allowing annotators to navigate the frames faster and more easily than before. Visible rolesets can be filtered according to the projects specified in the <usage> tags.

[16] Although it is very likely that the span will align with a node in the parse tree of a given sentence.

3 Fresh Corpora—New Domains and Genres

Several diverse corpora have been PropBanked and are now available on our GitHub site.

OntoNotes The first major expansion to the original WSJ PropBank was OntoNotes[17], described above.

BOLT The BOLT corpus (Garland et al., 2012; Song et al., 2014) was treebanked and PropBanked as part of the DARPA BOLT project. It is composed of 628,000 tokens of informal text, divided into SMS and text chat data (SMS), online discussion forums(DF), and translations of informal Arabic and Chinese data in English (CTS).

English Web Text The third corpus, the English Web Treebank (Bies et al., 2012), is 250k tokens of web text covering weblogs, newsgroups, emails, reviews and online question-answer pairs, and was funded by Google.

These three corpora not only provide three different genres, but each contains a wide range of subcorpora. One simple illustration of this within-corpus variety can be witnessed in the fact that the conversational speech in OntoNotes and in BOLT range from 7-10 words per sentence, whereas the OntoNotes weblog and BOLT discussion forums have an average sentence length of 20 words. One can see that each corpus contains very reduced, conversational examples such as the SMS, Emails, or the OntoNotes telephone conversation data. Similarly, each contains long, syntactically complex data—with data such as the BOLT Discussion Forum data differing from traditional newswire, not in complexity, but in editing and syntactic coherence.

3.1 Additional Diversification

Brown The original CoNLL-2005 task evaluated upon a small set of less than a thousand annotations. This corpus was augmented with additional annotation of some 15,000 verb predicates since the original CoNLL-2005 shared task. This larger dataset had preliminary analyses in (Pradhan et al., 2008), but was not released publicly. The updated version of this new corpus will be part of this collection. As one can see from Table 2, this annotation is entirely upon verbs, and therefore only measures verbal out-of-domain ability of models. Moreover, it should be noted that the Brown corpus—well-edited fiction texts released before 1961—depicts a very specific kind of out-of-domain test, and should likely be viewed as reflecting only one kind of out-of-domain performance.

LORELEI The English Reflex Core from DARPA LORELEI (Strassel and Tracey, 2016) consists of newswire text, a phrasebook, and an elicitation corpus. Approximately 100k English tokens (24k predicates) were manually treebanked and annotated with SRL. These sentences were also translated into twenty-four other languages to provide a parallel corpus for multi-lingual research.

Flickr-8k consists of image captions of the Flickr-8k corpus (Hodosh et al., 2013). The first large-scale PropBank project mapped to dependency trees involved the addition of SRL labels to Flickr image captions. 5147 image captions were double annotated and adjudicated. A first pass of annotation was completed on flat, unparsed sentences, followed by mappings to dependency parses.

ClearEarth The ClearEarth (Duerr et al., 2015, 2016) project aimed to port NLP tools to the earth sciences. This project produced annotated SRL corpora in several domains: sea ice blogs/news, sea ice academic journal articles, educational wiki on ecology (77k tokens), and earthquake (40k tokens). Both of these corpora will be released in the near future. Portions of the THYME corpus featured as data for TempEval shared tasks (Bethard et al., 2017). THYME corpus will be available soon on hNLP[18]

4 New Benchmarks

4.1 Evaluation Setup

The current, most common benchmarks for SRL comprise the OntoNotes v5.0 corpus (Pradhan et al., 2013; Weischedel et al., 2011) and a much smaller subset of the Brown corpus (and also the original WSJ subset with verb specific, and legacy annotations based on the first release of PropBank 1.0). These additional subcorpora, updated to match the revised, unified annotation guidelines and with a more generalized view of the concept of a *predicate* (i.e., including nouns and adjectives), can now supplant the common benchmarks for evaluations and provide a better view of the generalization capabilities of the latest SRL models.

[17] https://ontonotes.org

[18] https://healthnlp.hms.harvard.edu/center

Corpora	Genre	Predicate Type			
		Verbs (V)	Nouns (N)	(Light V)	Adj.
OntoNotes (ON)	NW, BN, BC, WB, TC, PT	349,352	40,163	(2,215)	750
English Web TB (EWT)	WB, QS	44,736	9,453	(732)	3,305
BOLT	CTS, SMS, DF	132,642	18,839	(1,973)	10,957
BROWN	FICTION, LETTERS, ETC.	15,646	0	(0)	0
LORELEI	WB	18,871	4,089	(196)	780
Flickr-8k	IMAGE CAPTIONS	5,897	551	(91)	51
ClearEarth	EARTH SCIENCES	10,070	5713	(8)	468
SHARP (hNLP)	CLINICAL NOTES	27,667	15,807	(22)	0
THYME (hNLP)	CLINICAL NOTES	49,649	17,906	(89)	756

Table 2: Core Corpora Annotated with PropBank rolesets for general English. Light verbs are annotated using nominal frames (Hwang et al., 2010) and therefore a subset of the nominal predicates.
Legends: NW: Newswire; BN: Broadcast News; BC: Broadcast Conversation; TC, CTS: Telephone Conversations; SMS: Text Messages; DF: Discussion forums; WB: Miscellaneous webdata; TB: Treebank

4.2 Choice of Tagger

We provide preliminary results on the performance of a state of the art, deep learning based tagger (Li et al., 2020) trained on the OntoNotes training data (Pradhan et al., 2013) which does not rely on an explicit syntactic structure. For the purposes of generating a baseline, neither did we retrain the model nor updated the constraints—rolesets, and other constraints—it uses during its structural tuning process.

4.3 Experiment Partitions

We reused all experimental partitions that were previously identified and used by other researchers. The two main examples of these are the CoNLL-2012 partitions[19] for the OntoNotes corpus and the Universal Dependencies (UD) partitions of the EWT and the Brown partitions that conform to the CoNLL-2005 evaluation and the experiments reported by Pradhan et al. (2008). We created new partitions for the BOLT data with an aim at stratification of the various sources and genres. All these partitions are explicitly available with the data and we plan to further ease their use by creating subdirectories within the git repository similar to the CoNLL-2012 partitions.

4.4 Recreating the Setup

As mentioned earlier, all the annotations will be available for download on the PropBank GitHub organization. All the annotations, except for the clinical notes and the earth sciences data will be made available as skeleton files exactly as in the case of the CoNLL-2012 release. Most of the underlying source text cannot be re-distributed owing to various copyright restrictions and needs to be obtained from LDC. The source text is present as part of the relevant corpora releases from LDC. The final evaluation data files can be created using the scripts provided on the git repository to populate the skeleton files with the words from the corpora releases by specifying the location of the downloaded corpora in the appropriate configuration files. Further details will be available in the documentation with the released corpora.

This mode of corpus distribution, though somewhat complex, has the advantage of making updated annotations available to the research community without having to make a separate release through LDC, which is not an instantaneous process. The underlying source text is not expected to change. It is well known that manually annotated data can never be perfect. There are always some errors that are found when the corpus is used by many researchers. Updating corpora too frequently to fix data errors has a negative effect of somewhat destabilizing the benchmarks and potentially obfuscating the interpretation of results. As a rule of thumb, releasing a new version of a corpus after a reasonable period of time (at least several years) allows the data to be cleaned of the inconsistencies[20]. This approach also allows a better workflow for incorporating corrections into the annotations when identified by the community via established

[19] https://github.com/ontonotes/conll-formatted-ontonotes-5.0

[20] This trend could be changing as better tools and evaluation infrastcutures become widely available.

software engineering best practices such as pull requests.

4.5 Regarding Conversational Data

One aspect of conversational data is the presence of noise in the form of restarts, repairs, disfluencies, non-speech words (laugh, cough, etc.). The Treebank annotations label conversational disfluencies and repairs with a specific EDITED phrase label, "He went (EDITED to) , to the store". The release in CoNLL-2012 removed those phrases from the surface strings for two main reasons: i) so that one could train upon the cleaned "He went to the store" instead; and ii) the coreference annotation ignored such cases anyway. The unified PropBank release follows the same approach for consistency. Though, given the reduced or eliminated reliance on parse structure for tagging semantic roles, it would be interesting to see if these artifacts can be learned and ignored by the deep learning models.

4.6 Experimental Results

The baseline results on the test partitions of four corpora are shown in Table 3 below. We use the CoNLL-2012 test set which is derived from the OntoNotes v5.0 corpus for evaluation and on which the semantic role labeling system has been trained. Note that there are two versions of the OntoNotes data. The second one uses the version of PropBank frame files that is consistent with the AMR frames. Notice that the inclusion of additional predicative parts of speech and more diverse genres increases the difficulty of the task significantly.

	Trained on OntoNotes v5.0 CoNLL-2012)
Test Set	F_1
ON (v5.0/CoNLL-2012)	86.7
ON (PB v3.4.0)	83.2
BOLT	80.1
EWT	80.5
BROWN	77.3

Table 3: Baseline performance on four main corpora annotated with PropBank v3.4.0 rolesets for English. The results include performance across all parts of speech. Follow latest updates and analysis at https://leaderboard.propbank.org

5 Summary and Discussion

This paper summarized the last twenty years of development and evolution of an approach to semantic role labeling called PropBanking. We've outlined the methods for converting PropBank to a unified form, and the advantages provided by that unified form and by the larger size of the PropBank corpora now available. The result is a set of consistently annotated corpora representing diverse genres and domains, all relying on a general set of English Frame Files. Where domain specific frame files are used, they are clearly marked. Tools are now available to view the frame files as a whole or as domain-specific subsets on an easily accessible web site. Similarly annotated corpora in several other languages are also available.

These new datasets offer opportunities for additional testing and evaluation that can advance the ability of SRL systems to generalize to new application areas and to new languages. We suggest that testing against the combination of OntoNotes, English Web Treebank, and BOLT corpora presented here can provide a more challenging SRL evaluation, requiring systems to better handle diverse domains and genres and non-verbal predicates.

In the coming year we look forward to toasting both PropBank on its 21st birthday and the winning systems of new SRL evaluation tasks.

References

Alan Akbik, Laura Chiticariu, Marina Danilevsky, Yunyao Li, Shivakumar Vaithyanathan, and Huaiyu Zhu. 2015. Generating high quality Proposition Banks for multilingual semantic role labeling. In *Proceedings of the 53rd Annual Meeting of the ACL and the 7th International Joint Conference on NLP*, pages 397–407, Beijing, China.

Daniel Albright, Arrick Lanfranchi, Anwen Fredriksen, William F Styler IV, Colin Warner, Jena D Hwang, Jinho D Choi, Dmitriy Dligach, Rodney D Nielsen, James Martin, et al. 2013. Towards comprehensive syntactic and semantic annotations of the clinical narrative. *Journal of the American Medical Informatics Association*, 20(5):922–930.

Olga Babko-Malaya, Ann Bies, Ann Taylor, Szu-ting Yi, Martha Palmer, Mitch Marcus, Seth Kulick, and Libin Shen. 2006. Issues in synchronizing the english treebank and propbank. In *Proceedings of the Workshop on Frontiers in Linguistically Annotated Corpora 2006*, pages 70–77.

Collin F. Baker, Charles J. Fillmore, and John B. Lowe. 1998. The Berkeley FrameNet project. In *Proceedings of the 17th International Conference on Compu-*

tational Linguistics (COLING/ACL-98), pages 86–90, Montreal.

Laura Banarescu, Claire Bonial, Shu Cai, Madalina Georgescu, Kira Griffitt, Ulf Hermjakob, Kevin Knight, Philipp Koehn, Martha Palmer, and Nathan Schneider. 2013. Abstract meaning representation for sembanking. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*.

Steven Bethard, Guergana Savova, Martha Palmer, and James Pustejovsky. 2017. SemEval-2017 task 12: Clinical TempEval. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 565–572, Vancouver, Canada. Association for Computational Linguistics.

Rajesh Bhatt, Bhuvana Narasimhan, Martha Palmer, Owen Rambow, Dipti Misra Sharma, and Fei Xia. 2009. A multi-representational and multi-layered Treebank for Hindi/Urdu. In *Proceedings of the Third Linguistic Annotation Workshop*, pages 186–189. Association for Computational Linguistics.

Ann Bies, Justin Mott, Colin Warner, and Seth Kulick. 2012. English Web Treebank (LDC2012T13). *Linguistic Data Consortium, Philadelphia, PA*.

Claire Bonial, Julia Bonn, Kathryn Conger, Jena Hwang, and Martha Palmer. 2014. PropBank: Semantics of New Predicate Types. In *LREC*, pages 3013–3019.

Claire Bonial, Julia Bonn, Kathryn Conger, and Jena D. Hwang. 2012. English propbank annotation guidelines.

Claire Bonial, Kathryn Conger, Jena D Hwang, Aous Mansouri, Yahya Aseri, Julia Bonn, Timothy O'Gorman, and Martha Palmer. 2017. Current directions in English and Arabic PropBank. In *Handbook of Linguistic Annotation*, pages 737–769. Springer.

Julia Bonn, Martha Palmer, Zheng Cai, and Kristin Wright-Bettner. 2020. Spatial AMR: Expanded spatial annotation in the context of a grounded Minecraft corpus. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 4883–4892, Marseille, France. European Language Resources Association.

Xavier Carreras and Lluís Màrquez. 2005. Introduction to the CoNLL-2005 shared task: Semantic role labeling. In *Proceedings of the Ninth Conference on Computational Natural Language Learning*, pages 152–164. Association for Computational Linguistics.

David Dowty. 1991. Thematic proto-roles and argument selection. *language*, pages 547–619.

R Duerr, A Thessen, CJ Jenkins, M Palmer, S Myers, and S Ramdeen. 2016. The ClearEarth Project: Preliminary findings from experiments in applying the CLEARTK NLP pipeline and annotation tools developed for biomedicine to the earth sciences. In *AGU Fall Meeting Abstracts*.

Ruth Duerr, Skatje Myers, Martha Palmer, Chris J Jenkins, Anne Thessen, and James Martin. 2015. Natural language processing and machine learning (nlp/ml): Applying advances in biomedicine to the earth sciences. In *AGU Fall Meeting Abstracts*, volume 2015, pages IN51A–1784.

Magali Sanches Duran and Sandra Maria Aluísio. 2011. Propbank-br: a brazilian portuguese corpus annotated with semantic role labels. In *Proceedings of the 8th Symposium in Information and Human Language Technology, Cuiabá/MT, Brazil*.

Christiane Fellbaum. 2010. WordNet. In *Theory and applications of ontology: computer applications*, pages 231–243. Springer.

Jennifer Garland, Stephanie Strassel, Safa Ismael, Zhiyi Song, and Haejoong Lee. 2012. Linguistic resources for genre-independent language technologies: user-generated content in BOLT. In *Workshop Programme*, page 34.

John J Godfrey, Edward C Holliman, and Jane McDaniel. 1992. Switchboard: Telephone speech corpus for research and development. In *Acoustics, Speech, and Signal Processing, 1992. ICASSP-92., 1992 IEEE International Conference on*, volume 1, pages 517–520. IEEE.

Jens Van Gysel, Meagan Vigus, Jayeol Chun, Kenneth Lai, Sarah Moeller, Jiarui Yao, Timothy J. O'Gorman, Andrew Cowell, W. Bruce Croft, Chu-Ren Huang, Jan Hajic, James H. Martin, Stephan Oepen, Martha Palmer, James Pustejovsky, Rosa Vallejos, and Nianwen Xue. 2021. Designing a uniform meaning representation for natural language processing. *Künstliche Intell.*, 35:343–360.

Katri Haverinen, Veronika Laippala, Samuel Kohonen, Anna Missilä, Jenna Nyblom, Stina Ojala, Timo Viljanen, Tapio Salakoski, and Filip Ginter. 2013. Towards a dependency-based PropBank of general Finnish. In *Proceedings of the 19th Nordic Conference of Computational Linguistics (NODALIDA 2013); May 22-24; 2013; Oslo University; Norway. NEALT Proceedings Series 16*, 085, pages 41–57. Linköping University Electronic Press.

Micah Hodosh, Peter Young, and Julia Hockenmaier. 2013. Framing image description as a ranking task: Data, models and evaluation metrics. *Journal of Artificial Intelligence Research*, 47:853–899.

Jena D Hwang, Archna Bhatia, Clare Bonial, Aous Mansouri, Ashwini Vaidya, Nianwen Xue, and Martha Palmer. 2010. PropBank annotation of multilingual light verb constructions. In *Proceedings of the Fourth Linguistic Annotation Workshop*, pages 82–90. Association for Computational Linguistics.

Ishan Jindal, Alexandre Rademaker, Michał Ulewicz, Huyen Nguyen, Ha Linh, Khoi-Nguyen Tran, Huaiyu Zhu, and Yunyao Li. 2022. Universal proposition bank 2.0. Marseille, France.

Paul Kingsbury and Martha Palmer. 2002. From Treebank to PropBank. In *LREC*, pages 1989–1993.

Tao Li, Parth Anand Jawale, Martha Palmer, and Vivek Srikumar. 2020. Structured tuning for semantic role labeling. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8402–8412, Online.

Adam Meyers, Ruth Reeves, Catherine Macleod, Rachel Szekely, Veronika Zielinska, Brian Young, and Ralph Grishman. 2004. The NomBank project: An interim report. In *HLT-NAACL 2004 workshop: Frontiers in corpus annotation*, volume 24, page 31.

Azadeh Mirzaei and Amirsaeid Moloodi. 2016. Persian Proposition Bank. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 3828–3835, Portorož, Slovenia.

Sarah Moeller, Irina Wagner, Martha Palmer, Kathryn Conger, and Skatje Myers. 2020. The Russian PropBank. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 5995–6002.

Anjali Narayan-Chen, Prashant Jayannavar, and Julia Hockenmaier. 2019. Collaborative dialogue in Minecraft. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5405–5415, Florence, Italy.

Tim O'Gorman, Sameer Pradhan, Martha Palmer, Julia Bonn, Kathryn Conger, and James Gung. 2018. The New Propbank: Aligning Propbank with AMR through POS Unification. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan.

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

Martha Palmer, Shijong Ryu, Jinyoung Choi, Sinwon Yoon, and Yeongmi Jeon. 2006. Korean PropBank. *LDC Catalog No.: LDC2006T03 ISBN*, pages 1–58563.

Martha Palmer, Nianwen Xue, Olga Babko-Malaya, Jinying Chen, and Benjamin Snyder. 2005b. A parallel Proposition Bank II for Chinese and English. In *Proceedings of the Workshop on Frontiers in Corpus Annotations II: Pie in the Sky*, pages 61–67. Association for Computational Linguistics.

Sameer Pradhan, Kadri Hacioglu, Valerie Krugler, Wayne Ward, James Martin, and Dan Jurafsky. 2005. Support vector learning for semantic argument classification. *Machine Learning Journal*, 60(1):11–39.

Sameer Pradhan, Alessandro Moschitti, Nianwen Xue, Hwee Tou Ng, Anders Björkelund, Olga Uryupina, Yuchen Zhang, and Zhi Zhong. 2013. Towards robust linguistic analysis using OntoNotes. In *Proceedings of the Seventeenth Conference on Computational Natural Language Learning*, pages 143–152, Sofia, Bulgaria.

Sameer S Pradhan, Wayne Ward, and James H Martin. 2008. Towards robust semantic role labeling. *Computational linguistics*, 34(2):289–310.

GG Sahin. 2016. Verb sense annotation for Turkish PropBank via crowdsourcing. In *Proceedings of 17th international conference on intelligent text processing and computational linguistics. CICLING*.

Karin Kipper Schuler. 2005. *VerbNet: A broad-coverage, comprehensive verb lexicon*. University of Pennsylvania.

Zhiyi Song, Stephanie Strassel, Haejoong Lee, Kevin Walker, Jonathan Wright, Jennifer Garland, Dana Fore, Brian Gainor, Preston Cabe, Thomas Thomas, et al. 2014. Collecting natural SMS and chat conversations in multiple languages: The BOLT phase 2 corpus. In *LREC*, pages 1699–1704.

Stephanie Strassel and Jennifer Tracey. 2016. Lorelei language packs: Data, tools, and resources for technology development in low resource languages. In *LREC*.

William F Styler, Steven Bethard, Sean Finan, Martha Palmer, Sameer Pradhan, Piet C De Groen, Brad Erickson, Timothy Miller, Chen Lin, Guergana Savova, et al. 2014. Temporal annotation in the clinical domain. *Transactions of the association for computational linguistics*, 2:143–154.

Mihai Surdeanu, Richard Johansson, Adam Meyers, Lluís Màrquez, and Joakim Nivre. 2008. The CoNLL-2008 shared task on joint parsing of syntactic and semantic dependencies. In *Proceedings of the Twelfth Conference on Computational Natural Language Learning*, pages 159–177.

Ann Taylor. 1996. Bracketing Switchboard: an addendum to the Treebank II bracketing guidelines. *Linguistic Data Consortium*.

Ralph Weischedel, Eduard Hovy, Mitchell Marcus, Martha Palmer, Robert Belvin, Sameer Pradhan, Lance Ramshaw, and Nianwen Xue. 2011. OntoNotes: A large training corpus for enhanced processing. In Joseph Olive, Caitlin Christianson, and John McCary, editors, *Handbook of Natural Language Processing and Machine Translation: DARPA Global Autonomous Language Exploitation*. Springer.

Kristin Wright-Bettner, Martha Palmer, Guergana Savova, Piet de Groen, and Timothy Miller. 2019. Cross-document coreference: An approach to capturing coreference without context. In *Proceedings of the Tenth International Workshop on Health Text Mining and Information Analysis (LOUHI 2019)*, pages 1–10, Hong Kong. Association for Computational Linguistics.