SemLink 2: Chasing Lexical Resources

Kevin Stowe[1], **Jenette Preciado**[2], **Kathryn Conger**[3], **Susan Brown**[3], **Ghazaleh Kazeminejad**[3], **James Gung**[3], **Martha Palmer**[3]
[1]Ubiquitous Knowledge Processing Lab (UKP Lab), Department of Computer Science
Technical University of Darmstadt
[2]SoundHound, Boulder, Colorado
[3]University of Colorado, Boulder
[1]stowe@ukp.informatik.tu-darmstadt.de
[2]jenette.preciado@gmail.com
[3]{firstname.lastname}@colorado.edu

Abstract

The SemLink resource provides mappings between a variety of lexical semantic ontologies, each with their strengths and weaknesses. To take advantage of these differences, the ability to move between resources is essential. This work describes advances made to improve the usability of the SemLink resource: the automatic addition of new instances and mappings, manual corrections, sense-based vectors and collocation information, and architecture built to automatically update the resource when versions of the underlying resources change. These updates improve coverage, provide new tools to leverage the capabilities of these resources, and facilitate seamless updates, ensuring the consistency and applicability of these mappings in the future.[1]

1 Introduction

Hand-crafted lexical resources remain an important factor in natural language processing research, as they can offer linguistic insights that are currently not captured even by modern deep learning techniques. SemLink is a connecting point between a number of different lexical semantic resources, providing mappings between different word senses and semantic roles, as well as a corpus of annotation (Palmer, 2009). SemLink has a variety of applications, from performing linguistic analysis of its component parts and their relations (Reisinger et al., 2015), extracting thematic role hierarchies (Kuznetsov and Gurevych, 2018), probing of linguistic formalisms (Kuznetsov and Gurevych, 2020), and computational methods for automatic extraction, improvement, and classification of computational lexical resources (Kawahara et al., 2014; Peterson et al., 2016, 2020).

[1]https://github.com/cu-clear/semlink

SemLink incorporates four different lexical resources: PropBank (Palmer and Kingsbury, 2005), VerbNet (Kipper-Schuler, 2005), FrameNet (Baker and Lowe, 1998), and WordNet via the OntoNotes sense groupings (Weischedel et al., 2011).[2] Each resource has different goals and benefits: WordNet has the greatest coverage, with very fine-grained word senses grouped into small "synonym sets". These are linked to each other with semantic relations like hyponymy and troponymy. PropBank defines the argument roles for its verb and eventive noun senses, information not available in WN. FrameNet groups verbs, eventive nouns and some adjectives into semantic frames, with fine-grained argument roles defined for each frame. These frames are linked by various relations, such as "inherited by" and "used by". VerbNet groups verbs into more or less semantically coherent classes based on shared syntactic alternations. This resource uses fairly coarse-grained argument roles and provides a list of typical syntactic patterns that the verbs of a class prefer. In addition, VN provides a semantic representation for each syntactic frame, using the class's argument roles in a first-order-logic representation that incorporates Generative Lexicon subevent structure.

Semlink provides a bridge between these resources, allowing users to take advantage of their different features and strengths. For example, the mappings between the semantic role labels allow users to accurately convert annotations done with PB roles to VN roles and combine their respective data sets into a much larger corpus of training and test data.

The goal of SemLink is to link senses between resources, maximizing the effectiveness of each. It is composed of two primary assets: mappings

[2]For the remainder of this work, we will refer to each by its acronym: PB, VN, FN, and ON, respectively.

Proceedings of the 14th International Conference on Computational Semantics, pages 222–227
June 17–18, 2021.

between resources, and a corpus of annotated instances. These are verbs in context that receive a PB roleset annotation, and VN class tag, a FN frame tag, and a sense tag based on the ON groupings.

The problem we address here is the constantly changing nature of these resources. They are evolving: new versions incorporate new semantics, new senses, better lexical coverage, and more consistent formatting. This makes it difficult to provide static links between them. SemLink has seen previous updates (Bonial et al., 2013) that improve consistency, but since that time many of the resources it links have undergone significant overhauls. Our work updates SemLink via four distinct contributions:

1. Automatic and manual updates to SemLink mappings based on new resource versions
2. Automatic addition of SemLink annotation instances, nearly doubling its size
3. Addition of sense embeddings and subject/object information
4. Release of software for automatic updates

2 Resources

A brief description of each resource in SemLink follows, along with the changes in each that have been implemented since the previous update.

2.1 PropBank

The previous version of SemLink incorporated PB annotation in the form of roleset mappings to VN classes and FN frames. It also contains gold annotation over sections of the Wall Street Journal corpus, with verbs annotated with their PB roleset. Each verb's arguments are annotated with their correct PB argument relations. These PB rolesets, mappings, and annotations remain core elements of SemLink, and we have expanded and updated each component for SemLink 2.0.

2.2 VerbNet

SemLink incorporates VN as an intermediary between the coarse-grained PB and fine-grained FN. Mapping files are provided that link PB rolesets to VN senses, which are then in turn linked to FN frames. The previous version of SemLink was built upon VN 3.2: this resource has since been updated to a new version (3.3), with substantial changes in class membership, thematic roles (Bonial et al., 2011), and semantics (Brown et al., 2018, 2019). We have incorporated these changes into SemLink 2.0 automatically where possible and manually where necessary.

2.3 FrameNet

The previous version of SemLink was built upon FN version 1.5; since then FN has released a new version (1.7), and this led to many consistency errors across resources. SemLink 2.0 provides manual updates to match the newest version of FN, as well as other consistency improvements.

2.4 OntoNotes Sense Groupings

The SemLink resource focuses less on these groupings than on PB, VN, and FN: it only includes ON as annotations on the provided instances. The ON resource has remained consistent since the release of the previous SemLink version, and thus the instance annotations remain valid.

3 Improvements and Additions

SemLink incorporates these resources via mapping files (for PB, VN, and FN) and predicate instance annotations (including all four resources). We will now overview each of these artifacts, highlighting the updates in our new release and the tools and practices used to generate these updates.

3.1 PB to VN mappings

The previous version of SemLink contains two files comprising the mappings from PB to VN: a mapping file that links PB rolesets to VN senses, and a mapping file linking PB arguments (ARG0, ARG1, etc) to VN thematic roles (Agent, Patient, etc). These files contain a growing number of inaccuracies as the resources have been updated, particularly with PB's update to unified frame files and VN's update to the version 3.3.

To deal with these constant updates, we've improved the system that automatically generates these mapping files based on ground-truth mappings present in PB. The PB frame files contain links from each roleset to possible VN classes: this allowed us to generate a large number of accurate mappings based purely on the information present in PB. The main update to this architecture is the development of VN class matching. We can now find if verbs have moved between classes, allowing the automated updater to find more valid instances. This system incorporates soft class matching for when verbs moved between VN subclasses, as well as exploiting available WordNet mappings in VN to identify if a verb moved to a new class.

The mappings generated by this system are not exhaustive: the ever-changing nature of the two projects makes it impossible to have all possible mappings. One of the primary goals of SemLink is to ensure that the most consistent possible mappings between resources is available, and our update helps to foster this consistency by making available our software for updating and evaluating the accuracy of these mappings. This is done by automatically generating mappings from PB to VN based on PB frame files, combining them with the previous version of manual mappings, and checking both of these mappings for consistency.

This process produces an update mapping resource from PB to VN. While these mappings don't eliminate the need for some manual annotation, as substantive changes can require new mappings to be added or deleted, it does allow the resource to be consistently and automatically updated while preserving only valid mappings.

3.2 VN to FN mappings

SemLink contains similar mapping files from VN to FN: one mapping from VN senses to FN frames, and one mapping from VN thematic roles to FN's typically more specific frame elements. As with PB and VN, FN has seen a significant update (to version 1.7) since the previous SemLink release, and these mappings files have become outdated.

Unlike PB, neither VN nor FN implicitly keeps track of mappings to the other resource: the only linking between them is in SemLink's mapping files. Therefore, for these files, we employed a semi-automated system to identify incorrect mappings and make updates. We run a script to identify whether VN class/role and FN frame/frame elements are valid. This is done by checking if the classes, roles, frames and frame elements still exist in the current version of the resource, and then checking if the roles and frame elements are still valid for the given classes and frames. We then pass them to annotators if there are errors. This was done for all of the mappings in the previous version, yielding 2,387 valid mappings, 160 of which came from manual re-annotation. These mappings were then compiled to form the new VN to FN mapping file for SemLink 2.0.

For both PB to VN and VN to FN mappings, we employed automatic procedures that allowed us to update outdated SemLink instances to match the current resources. However, these updates are necessarily not comprehensive: we only updated instances for which we could identify automatic mappings between old and new. If the resources changed in unpredictable ways (ie. a sense tag changed itself changed meanings), these mappings may still be inconsistent. We therefore include for each instance in SemLink 2.0 and indicator for each mapping whether it was derived from an automatic procedure or manually annotated.

Previous Version		SemLink 2.0		
Resource	Count	Count	Added	Coverage
PB	75k	148k	73k	.99
VN	75k	97k	22k	.65
FN	37k	42k	5k	.28
ON	28k	48k	21k	.33
Total	75k	149k	74k	+98%

Table 1: Summary of Annotation Updates to SemLink

3.3 Annotations

The second artifact produced for SemLink is a set of annotations. These consist of predicates annotated with PB frames, VN senses, FN frames, ON groupings, and each resource's representation of the predicates' arguments. An example of an annotation instance is shown in Figure 1.

3.3.1 Updates to Previous Annotations

All instances underwent an automatic update process based on our revision of mapping resources. The sense tags for each resource are validated, and automatically updated via mappings if errors are found. This process is repeated for role arguments.

This was done for the 74,920 instances available with the previous SemLink. In order to keep the resource as large and as flexible as possible, as long as an instance had a PB roleset, we didn't remove instances with invalid mappings: rather, we kept these instances and left the additional information (VN, FN, etc) as "None". This allows us to maintain the size of the resource and while preserving only the accurate annotations.

3.3.2 New Annotations

In addition to updating the previous annotations, we were also able to leverage additional annotation projects to expand the scope of the SemLink resource. We gathered 72,822 additional instances from the OntoNotes 5.0 release annotated with the unified PB rolesets (Weischedel et al., 2011), and employed our updated mapping files to automatically attribute VN and FN information to them. We also collected 5,300 instances that were manually

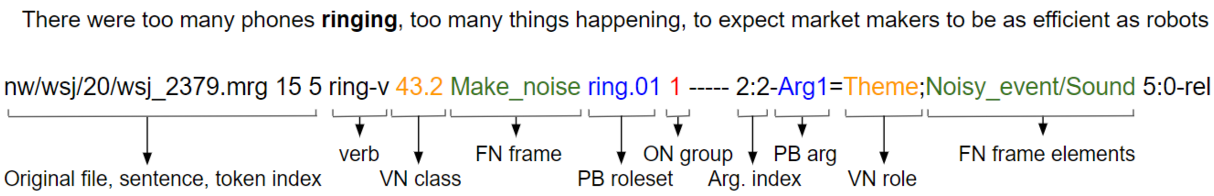

Figure 1: SemLink annotation instance for the verb "ringing" in the above sentence.

annotated with VN classes (Palmer et al., 2017), and extracted PB and FN information from these based on mapping files.

Similar to the updates above, we automatically check these instances to determine if their annotations were valid (the class, sense, or frame still exists) in the modern versions of each resource. and then added them to SemLink's annotation corpus. A summary of the update to the annotations is shown in Table 1.

From this summary we can see substantial improvements to the dataset across all resources, with the greatest impact coming from the new annotations. However, as we automatically add instances based on PB and VN annotation, they often lack mappings to the other resources. This, combined with the fact that some VN and FN annotations were removed due to inconsistency with the latest versions, leads to a decrease in the percent of instances tagged with each particular resource, despite the increase in total annotations.

3.4 VN Tools

In order to ensure the applicability of these mappings and lexical resources, we include two additional components: sense embeddings and common arguments. These are based on VN, as it directly links to PB and FN.

3.4.1 VN Embeddings

We train embeddings based on VN in a style similar to that of (Sikos and Padó, 2018). We tag a corpus of 4.5m sentences from Wikipedia with a VN class tagger (Palmer et al., 2017). We then learn embeddings for both VN classes and specific VN senses by modifying the resulting corpora. First, to generate generic VN class embeddings, we replace the verb directly with its labeled class. This allows the embedding model to learn a representation that generalizes over all instances of a particular VN class, and provides an abstraction away from the individual lexical items. Second, to generate sense-specific word embeddings, we concatenate the class information along with the verb. This yields more specific embeddings that concretely reflect contextual usages of the given verb. The resulting sentences can then be fed to a lexical embedding algorithm of choice: here we use GloVe (Pennington et al., 2014) and Word2Vec (Mikolov et al., 2013) embeddings of size 100.

These embeddings have proven an effective addition to traditional embeddings for classification tasks, and even have advantages over contextual embeddings. Stowe (2019) show that incorporating VN-based sense embeddings into LSTM-based metaphor detection improves results over using ELMo embeddings alone, despite the fact that the contextualized ELMo embeddings should independently capture sense information (Peters et al., 2018).[3]

These methods for learning embeddings are broadly applicable to any lexical resource, and are adaptable to changing versions; the embeddings provided are trained using VN 3.3, and as we provide links from VN to PB and FN, we further believe that the accompanying embeddings can be directly linked to these two resources.

3.4.2 VN Common Arguments

In addition to embeddings, we also collect argument information based on VN class tagging. We collect for each class the most frequent subjects and objects of verbs tagged with that class. This is done by tagging the above Wikipedia corpus with VN classes, then using a dependency parser to extract subject and object information (Chen and Manning, 2014). This automated procedure does inherently introduce noise, but it allows us to form a general idea of kind of arguments that typify the semantic roles and to better understand the syntactic and collocational properties of verb classes. Practitioners who are researching verb classes can use these to better understand from a quantitative perspective what kinds of subjects and objects are likely to ap-

[3] Note that these results are from embeddings trained on VN version 3.2; they have since been updated to version 3.3

pear with given verb classes, further facilitating research into lexical semantics.

3.5 Software

In order to manage these updates, we've built a substantial number of infrastructure components to support the interaction between these resources. This includes interfaces to each resource, to SemLink, and tools for making automatic updates based on different versions. The SemLink scripts have the flexibility to use and compare various different versions of each resource; this allows us to quickly update SemLink to new versions.

This software will be released along with the new version via GitHub, with the hope that the community can maintain and improve its functionality as necessary, and to allow researchers to be able to easily interact with both the resources linked and the SemLink resource itself. Critically, this resource will mitigate the damage of future changes to each individual resource, as SemLink can painlessly be updated to accommodate new versions.

4 Conclusions and Future Work

Our updates to SemLink consist of four main components. (1) We update SemLink data to match the current versions of each resource through automatic and manual methods. (2) We add annotations to improve the coverage of the resource. (3) We add sense embeddings and argument information. (4) We provide automatic tools to allow the SemLink resource to be consistently updated. As these lexical resources are always changing, these tools are necessary for the resource to remain viable, and while the process of linking semantic resources can likely never be fully automated, these tools can assist in this process. This work then comes with two artifacts: the new SemLink resource (mapping files and annotations) as well as architecture for updating and managing SemLink.

The coverage is by no means complete and many lexical items in each resource contain no viable mappings. Manual annotation of links between resources is essential for the success of the SemLink resource: while we can automatically filter out inaccurate mappings when resources change, this leaves blind spots where we have incomplete mappings, and manual annotation is currently the most accurate way to cover these gaps.

Another direction of future work is evaluating the usefulness of these linked resources. While there have been evaluations comparing the three semantic role labelling frameworks provided via PB, VN, and FN (Hartmann et al., 2017), a full-scale evaluation of the links between them is yet to be done, and may provide valuable insight not only into how to best improve SemLink, but also into how these kinds of linked resources can be best employed. While modern NLP focuses largely around end-to-end models that implicitly capture semantic relations, there is still a role for hand-curated lexical resources to play, and we believe SemLink can be an effective resource for those studying computational lexical semantics, word sense disambiguation and semantic role labelling, and other tasks requiring linked lexical resources.

5 Acknowledgements

We gratefully acknowledge the support of DTRAl-16-1-0002/Project 1553695, eTASC - Empirical Evidence for a Theoretical Approach to Semantic Components and DARPA 15-18-CwC-FP-032 Communicating with Computers, C3 Cognitively Coherent Human-Computer Communication (sub from UIUC) and Elementary Composable Ideas (ECI) Repository (sub from SIFT), and DARPA FA8750-18-2-0016-AIDA – RAMFIS: Representations of vectors and Abstract Meanings for Information Synthesis. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of DARPA, DTRA, or the U.S. government.

References

Fillmore C.J. Baker, C. F. and J.B. Lowe. 1998. The Berkeley FrameNet project. pages 86–90, Montreal, QC. COLING-ACL '98.

Claire Bonial, William Corvey, Martha Palmer, Volha V Petukhova, and Harry Bunt. 2011. A hierarchical unification of lirics and verbnet semantic roles. In *2011 IEEE Fifth International Conference on Semantic Computing*, pages 483–489. IEEE.

Claire Bonial, Kevin Stowe, and Martha Palmer. 2013. Renewing and revising SemLink. In *Proceedings of the 2nd Workshop on Linked Data in Linguistics (LDL-2013): Representing and linking lexicons, terminologies and other language data*, pages 9 – 17, Pisa, Italy. Association for Computational Linguistics.

Susan Windisch Brown, Julia Bonn, James Gung, Annie Zaenen, James Pustejovsky, and Martha Palmer. 2019. Verbnet representations: Subevent semantics for transfer verbs. In *Proceedings of the First International Workshop on Designing Meaning Representations*, pages 154–163.

Susan Windisch Brown, James Pustejovsky, Annie Zaenen, and Martha Palmer. 2018. Integrating generative lexicon event structures into verbnet. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC-2018)*.

Danqi Chen and Christopher Manning. 2014. A fast and accurate dependency parser using neural networks. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 740–750, Doha, Qatar. Association for Computational Linguistics.

Silvana Hartmann, Éva Mújdricza-Maydt, Ilia Kuznetsov, Iryna Gurevych, and Anette Frank. 2017. Assessing SRL frameworks with automatic training data expansion. In *Proceedings of the 11th Linguistic Annotation Workshop*, pages 115–121, Valencia, Spain. Association for Computational Linguistics.

Daisuke Kawahara, Daniel W. Peterson, and Martha Palmer. 2014. A step-wise usage-based method for inducing polysemy-aware verb classes. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1030–1040, Baltimore, Maryland. Association for Computational Linguistics.

K Kipper-Schuler. 2005. VerbNet: A broad-coverage, comprehensive verb lexicon.

Ilia Kuznetsov and Iryna Gurevych. 2018. Corpus-driven thematic hierarchy induction. In *Proceedings of the 22nd Conference on Computational Natural Language Learning*, pages 54–64, Brussels, Belgium. Association for Computational Linguistics.

Ilia Kuznetsov and Iryna Gurevych. 2020. A matter of framing: The impact of linguistic formalism on probing results. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 171–182, Online. Association for Computational Linguistics.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *CoRR*, abs/1301.3781.

Gildea D. Palmer, M. and P. Kingsbury. 2005. The Proposition Bank: An Annotated Corpus of Semantic Roles. volume 31, pages 71–106.

M. Palmer. 2009. Semlink: Linking PropBank, VerbNet and FrameNet. Pisa, Italy. Proceedings of the Generative Lexicon Conference.

Martha Palmer, James Gung, Claire Bonial, Jinho Choi, Orin Hargraves, Derek Palmer, and Kevin Stowe. 2017. The Pitfalls of Shortcuts: Tales from the word sense tagging trenches. *Essays in Lexical Semantics and Computational Lexicography - In honor of Adam Kilgarriff.*

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.

Daniel Peterson, Jordan Boyd-Graber, Martha Palmer, and Daisuke Kawahara. 2016. Leveraging VerbNet to build corpus-specific verb clusters. In *Proceedings of the Fifth Joint Conference on Lexical and Computational Semantics*, pages 102–107, Berlin, Germany. Association for Computational Linguistics.

Daniel Peterson, Susan Brown, and Martha Palmer. 2020. Verb class induction with partial supervision. In *Proceedings of the Thirty-fourth AAAI Conference on Artificial Intelligence*, New York City, NY.

Drew Reisinger, Rachel Rudinger, Francis Ferraro, Craig Harman, Kyle Rawlins, and Benjamin Van Durme. 2015. Semantic proto-roles. *Transactions of the Association for Computational Linguistics*, 3:475–488.

Jennifer Sikos and Sebastian Padó. 2018. Using embeddings to compare FrameNet frames across languages. In *Proceedings of the First Workshop on Linguistic Resources for Natural Language Processing*, pages 91–101, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Kevin Stowe. 2019. Syntactic and semantic improvements to computational metaphor processing.

R. Weischedel, E. Hovy, M. Marcus, M. Palmer, R. Belvin, S. Pradan, L. Ramshaw, and X. Nianwen. 2011. OntoNotes: A Large Training Corpus for Enhanced Processing. *Handbook of Natural Language Processing and Machine Translation: Global Automatic Language Exploitation*, pages 53–63.