Leveraging Heterogeneous Data Sources for Relational Semantic Parsing

First Author
Affiliation / Address line 1
Affiliation / Address line 2
Affiliation / Address line 3
email@domain

Second Author
Affiliation / Address line 1
Affiliation / Address line 2
Affiliation / Address line 3
email@domain

Abstract

A number of semantic annotation efforts have produced a variety of annotated corpora, capturing various aspects of semantic knowledge in different formalisms. Due to to the cost of these annotation efforts and the relatively small amount of semantically annotated corpora, we argue it is advantageous to be able to leverage as much annotated data as possible. This work presents a preliminary exploration of the opportunities and challenges of learning semantic parsers from heterogeneous semantic annotation sources.

1 Introduction

Multiple annotated resources capture relational semantic information, including FrameNet (Fillmore et al., 2002), PropBank (Palmer et al., 2005), VerbNet (Kipper et al., 2008), and the AMR Bank (Banarescu et al., 2013) for English. Each of these is the result of painstaking lexicography and annotation efforts by teams of linguists over several years. However, they are the result of largely independent annotation efforts which make different theoretical commitments. Here, we consider how we might best leverage such heterogenous resources to improve semantic parsing, so as to let semantic resources created at considerable cost not go to waste.

Supervised semantic parsers are usually trained with the representation and annotations from a single resource. For instance, SEMAFOR (Das et al., 2010, 2013) is a frame-semantic parser trained on the full-text annotated corpus of FrameNet. PropBank has likewise served as the training resource for semantic role labeling (SRL) systems (e.g., Punyakanok et al., 2008; see Palmer et al., 2010 for a review). FrameNet and PropBank seek to encode very similar aspects of meaning, including event predicates and their labeled arguments; yet the two projects have developed different representations, lexicons, and corpus annotation approaches due to different design considerations. Broadly speaking, FrameNet contains richer forms of linguistic detail (e.g., predicates are semantically organized with respect to one other in FrameNet), but the PropBank representation is much more conducive to large-scale annotation, and PropBank therefore has much greater type and token coverage of verbs. Moreover, the annotated corpora of FrameNet and PropBank capture different domains. As such, the two resources can be seen as partially overlapping but partially complementary. We hypothesize that systems could exploit this complementarity to gain robustness at semantic analysis. Encouragingly, some studies of *syntactic* dependency parsing have demonstrated that divergent treebanks can be leveraged to improve parsing quality (Zhou and Zhao, 2013; Johansson, 2013). A motivating example appears in §2, followed by quantative analysis of the resources in §3.

The problem of combining the knowledge across these resources can be viewed as a transfer learning or joint learning problem. This problem also manifests itself in the information extraction community, where there are multiple knowledge bases conforming to different ontologies (e.g., Riedel et al., 2013). We consider several possible modeling frameworks below in §4.

2 Example

An example illustrating the need for models that integrate resources is as follows:

> **Taxing** wealthy individuals **appeals** to liberals; Brown **urges abolishing** regulations.

The four verbal predicates in this sentence are bolded. Figure 1 shows the frame-semantic parse

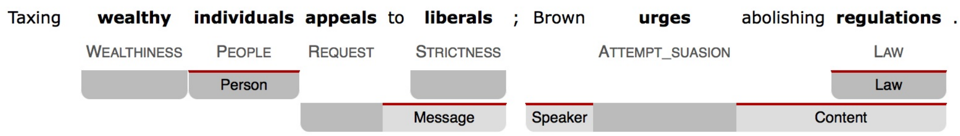

Figure 1: Frame semantic parse from SEMAFOR for a constructed example sentence.

in which SEMAFOR has predicted frame-evoking *targets* (predicate tokens), *frame labels* (displayed immediately below the targets), *argument spans* (red lines) for each frame, and *argument labels* (frame element names). We observe that:

1. **urges** is mapped correctly to a frame, and its arguments to frame elements.
2. **individuals** receives the correct frame label, PEOPLE, but is missing an argument ("wealthy" should fill the *Descriptor* frame element). In other wordings of this sentence, "wealthy" was correctly identified as an argument to PEOPLE, but incorrectly labeled with the *Ethnicity* frame element.
3. **appeals** is correctly identified as a target, but the wrong frame label is chosen (a sense disambiguation error). The parse suggests that somebody is making an appeal to liberals, whereas the correct analysis would use the EXPERIENCER_OBJ frame to represent that an idea (taxing wealthy individuals) provokes an emotional response on the part of liberals. Because the inventory of valid arguments is different for each frame, and SEMAFOR predicts frame labels first, an error in frame identification will likely produce incorrect argument analyses.
4. **abolishing** is not identified as a target because it is absent from the FrameNet lexicon. In principle, however, it should be recognized as evoking the PROHIBITING frame, which contains nearly synonymous verbs. SEMAFOR attempts to disambiguate such OOV lexical items to existing frames (Das and Smith, 2011).
5. **taxing** is not identified as a target because it, too, is absent from the FrameNet lexicon. However, there is no apparent home for it in any existing frame; a TAXATION frame would have to be defined in the lexicon.

All of these verbs are annotated in the PropBank corpus, which assigns a coarse sense number to each predicate and describes the predicate's core and non-core arguments. Even though PB and FN use different label spaces for predicates/frames and arguments/frame elements, we propose that PB could provide useful information for frame-semantic parsing with respect to (a) detecting and labeling arguments (#2 above), (b) disambiguating targets that have multiple senses in both the PB and FN lexicons (#3 above), and (c) labeling arguments where SEMAFOR is able to identify the correct frame for a target that is not in FN, provided that some other verbs in the same frame *are* present in both PB and FN (#4 above).

3 Quantitative analysis

Here we quantify the extent to which the PropBank (PB) and FrameNet (FN) resources do or do not overlap. We obtain verb coverage measurements from version 1.5 of FrameNet and the PropBank WSJ (PB-WSJ) annotations in version 5.0 of OntoNotes (**?**).

3.1 Lexical coverage

As noted above, type and token coverage is an advantage that PB has over FN, at least for verbs.

Type coverage. Overall, there are 1,260 verb types in the PB-WSJ data that are not present at all in the FN lexicon. These include **involve**, **lurk**, **nominate**, **ladle**, **entice**, and **bank**. This is partially due to missing frames, and partially due to coverage gaps within existing frames: e.g., the GIVING frame has 19 known verbs as targets in FN, but many plausible members of this frame—**allot**, **assign**, **designate**, **allocate**, etc.—are not present in any FN frame.

Corpus support for lexical types. 984 of 4,894 verb senses (frame-disambiguated lexical units) in the FN lexicon never occur in the full-text annotations; many more occur only a few times. For instance, the MANIPULATION sense of **hold** is unattested in the full-text annotations, whereas PB-WSJ has 177 instances (that had an extractable argument mapping) of the corresponding verb senses according to SemLink, **hold.01** and **hold.06**. The frame EXPERIENCER_OBJ has several predicates without any annotations, for example **appeal**, **harass**, **worry**, and **boggle**. The first three of these have PB-WSJ annotations.

3.2 SemLink

The SemLink database (Bonial et al., 2014) specifies mappings between PropBank, FrameNet, and

FN frame annotation	PB verb tokens	% of all
Frame label = `NF`	14,624	20%
Frame label = `IN`	22,982	31%
Frame with no arguments	15,533	21%
Frame with at least 1 mappable argument	15,323	20%
Instances not mapped due to other issues	6,516	9%
Total	74,977	100%

Table 1: Statistics of PB-WSJ data from SemLink

Annotation unit	Count
Sentences in FN 1.5	5,946
FN frame annotations for verb predicates	6,993
Verb annotation density	1.2
PB-WSJ sentences	35,426
PB-WSJ verb tokens	96,517
Verb annotation density	2.7
SL-WSJ sentences with at least one mappable FN frame and argument	12,382
Mapped frame annotations	15,323
Verb annotation density	1.2

Table 2: Comparison of annotation density

VerbNet. It contains *sense-level mappings* between VerbNet and each of the other resources; because the senses in these resources may represent different interpretations and granularities of concepts, the sense-level mappings may be one-to-one, one-to-many, or many-to-many. These mappings link PropBank and FrameNet indirectly via VerbNet. Second, SemLink provides some *token-level parallel annotations* for the 3 representations in a subset of the PB-WSJ text: hereafter SL-WSJ.

We focus on using token-level SemLink version 1.2.2c annotations as a (disambiguated) mapping from PB to FN tokens. Some statistics appear in tables 1 and 2. The first table summarizes the SemLink mappings in the SL-WSJ data. Of 74,977 SL-WSJ verbs, a majority cannot be mapped to FN labels for various reasons. Around 31% of the predicates have the frame label `IN` ("indefinite") where the mapping from VerbNet to FrameNet is ambiguous. About 20% of the instances are labeled `NF` ("no frame"), indicating a coverage gap in FrameNet. 21% of verbs have frame labels but no frame element annotations. Most of these are predicates with modifier arguments. Other arguments pointed to null anaphora that could not be resolved to overt arguments. This leaves 15,323 mappable instances with at least one overt argument, or 20% of SL-WSJ verbs.

The second table compares the extracted SL-WSJ annotations with the full-text FN annotations used to train SEMAFOR. It seems that the PB sentences contain a higher rate of annotated verbs per sentence than FN, likely due in part to gaps in coverage for FN.[1] There is some dropoff in annotation density when mapping from PB to FN due to coverage gaps in SemLink.

The mappings we obtained increase the number of annotations for around 170 frames. Figure 2 shows a stacked bar chart plotting the annotations for every frame, sorted in decreasing order of additional frames obtained. The highest is around 1,500 new annotations for the STATEMENT frame.

[1]Other possible explanations include domain differences and different conventions for light verbs.

Noise in SemLink. Some of the FN information in SemLink is out of date due to subsequent changes in FrameNet. For example, the frame STATEMENT no longer contains the lexical unit **complain.v**, which has been moved to a new frame (COMPLAINING). There are around 3,000 such instances with obsolete annotations. Some of them may be updated automatically using the Reframing_mapping pointers in FrameNet, but some may have to be reannotated. There are some erroneous FN annotations as well: e.g., all 14 instances of **liquidate** are labeled KILLING, despite being used in the financial sense; and in 17 cases **direct** is erroneously marked as BEHIND_THE_SCENES (i.e., film direction). Therefore, we suspect that heavy reliance on the SemLink annotations in a model will be a source of precision errors.

4 Modeling

This problem of learning from multiple resources can be formulated in different ways. We present here a spectrum of possible approaches. For simplicity, let us assume there are two resources, D_1 and D_2. Let $X_1, X_2 \in \mathcal{X}$ represent the set of data instances (i.e. sentences) from the two sources and $Y_1 \in \mathcal{Y}$, $Y_2 \in \mathcal{Y}'$ be the labels.

- **Deterministically unifying the schemas for the training data.** With a mapping $\phi : \mathcal{Y}' \rightarrow \mathcal{Y}$ between the label spaces from the two resources, the annotations from one resource can be transformed into the labeling schema used by the other, in order to train a single model on $\langle X_1, Y_1 \rangle \cup \langle X_2, \phi(Y_2) \rangle$. The (limited and ambiguous) SemLink mappings discussed above are one possible choice of ϕ; a latent mapping could also be learned, treating the noisy correspondences in SemLink as evidence.

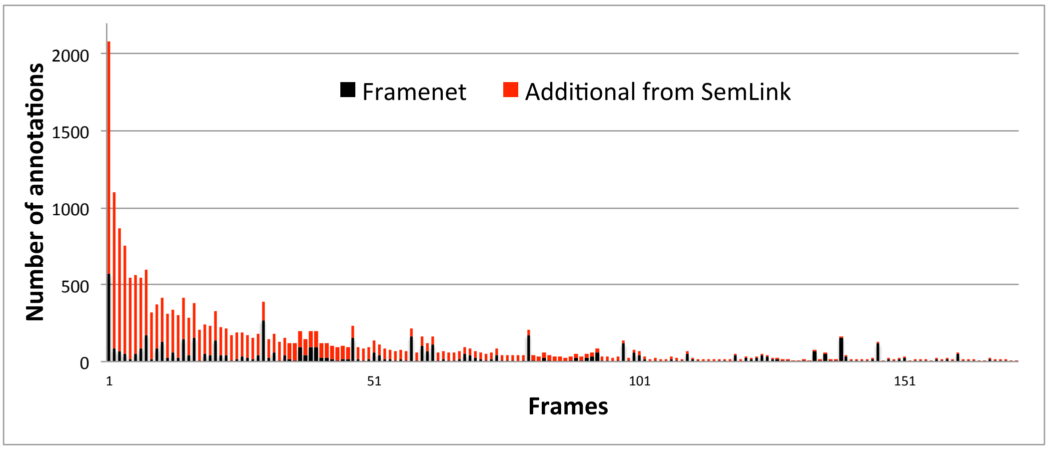

Figure 2: The number of new annotations per frame obtained upon extracting the WSJ mapping from SemLink. The bars are coloured (red) to indicate the contribution from the new annotations. Only frames for which new annotations were found are shown.

- **Learning and decoding as a pipeline.** Training a source model on D_1 and applying it on X_2 to extract features for training a target model on D_2. Johansson (2013) calls this *guided parsing*.
- **Bootstrapping** with approaches such as co-training (Blum and Mitchell, 1998; Clark et al., 2003), where two separate models are iteratively improved by providing high-confidence pseudo-annotated training examples for each other. A similar approach is taken in Zhou and Zhao (2013) for dependency parsing.
- **Optimizing a multi-task objective** with a loss function over the two data sources D_1 and D_2. The feature spaces from the two tasks can be combined (Daumé, 2007; Johansson, 2013) or coupled together using a feature transformation (Argyriou et al., 2006; Blitzer et al., 2006).

5 Conclusion

Based on our analysis, we can conclude that learning models from heterogeneous data sources is a promising direction for improving the performance of semantic parsing systems. With several learning frameworks at our disposal, we hope to develop an approach that compensates for gaps in both lexical and corpus coverage to improve performance, particularly for the FrameNet task.

References

A. Argyriou, T. Evgeniou, and M. Pontil. 2006. Multi-task feature learning. *NIPS*.

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 *Proc. of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pages 178–186.

John Blitzer, R. McDonald, and F. Pereira. 2006. Domain adaption with structural correspondence learning. *EMNLP*.

Avrim Blum and Tom Mitchell. 1998. Combining labeled and unlabeled data with co-training. In *Proc. of COLT*, pages 92–100. Madison, WI.

Claire Bonial, Kevin Stowe, and Martha Palmer. 2014. Renewing and revising semlink. *ACL*.

Stephen Clark, James R. Curran, and Miles Osborne. 2003. Bootstrapping POS taggers using unlabelled data. *CoNLL*.

Dipanjan Das, Desai Chen, André F. T. Martins, Nathan Schneider, and Noah A. Smith. 2013. Frame-Semantic Parsing. *Computational Linguistics*, 40(1):9–56.

Dipanjan Das, Nathan Schneider, Desai Chen, and Noah A. Smith. 2010. Probabilistic frame-semantic parsing. *NAACL-HLT*.

Dipanjan Das and Noah A. Smith. 2011. Semi-supervised frame-semantic parsing for unknown predicates. In *ACL*.

Hal Daumé, III. 2007. Frustratingly easy domain adaptation. In *Proc. of ACL*, pages 256–263. Prague, Czech Republic.

Charles J. Fillmore, Christoper R. Johnson, and Miriam R. L. Petruck. 2002. Background to framenet. *International Journal of Lexicography*.

Richard Johansson. 2013. Training parsers on incompatible treebanks. In *Proc. of NAACL-HLT*, pages 127–137. Atlanta, United States.

Karin Kipper, Anna Korhonen, Neville Ryant, and Martha Palmer. 2008. A large-scale classification of english verbs. *Language Resources and Evaluation Journal*.

Martha Palmer, Daniel Gildea, and Paul Kingsbury. 2005. The proposition bank: An annotated corpus of semantic roles. *Computational Linguistics*.

Martha Palmer, Daniel Gildea, and Nianwen Xue. 2010. Semantic Role Labeling. *Synthesis Lectures on Human Language Technologies*, 3(1):1–103.

V. Punyakanok, D. Roth, and W. Yih. 2008. The importance of syntactic parsing and inference in semantic role labeling. *Computational Linguistics*.

Sebastian Riedel, Limin Yao, Andrew McCallum, and Benjamin M. Marlin. 2013. Relation extraction with matrix factorization and universal schemas. In *Proc. of NAACL-HLT*, pages 74–84.

Guangyou Zhou and Jun Zhao. 2013. Joint inference for heterogeneous dependency parsing. In *ACL*.