Research Statement

Nathan Schneider
November 2015

1 Introduction

Imagine next-generation software that assists users learning to write in a nonnative language. For any topic, it helps the user to *say what they mean* clearly and fluently. Not only does it suggest corrections, it also explains its feedback, searches the web for similar examples, and offers practice exercises. This software does not exist today, but many of the necessary building blocks do. Among other things, it would require modules that robustly negotiate between intended meanings and the vocabulary and grammar of a language. The field of natural language processing (NLP) aims to assemble a common toolbox for automatic language analysis, whose tools could help power applications ranging from search engines to conversational agents to machine translation systems, as well as language-driven scholarship in the sciences and humanities.

I seek to "reverse-engineer" natural language from a linguistic and computational perspective. My research follows the life cycle of textual data as its linguistic structure and semantic content are analyzed by humans and machines. In particular, I design syntactic and semantic *abstractions* that are informed by linguistic theory, but are sufficiently formal, general-purpose, and scalable to suit statistical NLP—as demonstrated, in part, through human annotation of text corpora. In addition, I develop NLP algorithms and techniques to learn these novel abstractions from the annotated datasets. Thus, my work tightly integrates linguistic description, data annotation, and machine learning in the service of natural language understanding.

To make NLP systems robust, it is crucial to assemble ample data resources and, via statistical methods, to learn to generalize to new data. Scale is a fundamental concern here: not only algorithms and systems, but also annotation methods, should be efficient and should apply to a broad swath of language data.

In light of these concerns, I work to answer questions such as: What tools belong in the NLP toolbox? How can they be made flexible enough to handle all sorts of language varieties (e.g., social media text)? What linguistic theories can be brought to bear, and what data resources can be efficiently obtained, to make these tools possible?

2 Lexical Semantic Analysis for Domain-General Language Understanding

Analyzing word meanings with broad coverage. My dissertation developed a method of lexical semantic description for broad-coverage annotation and modeling in running text (Schneider, 2014). The approach circumvents limitations of traditional dictionary-based approaches, which are costly to build and annotate with; are not easily adapted to new vocabulary, genres, or languages; and are computationally impractical due to large label spaces. It segments sentences into **lexical expressions** and enriches some of those expressions with domain-independent semantic class labels called **supersenses** (Ciaramita and Altun, 2006). For instance, in the sentence

A	Junction City	chocolate lab	gave birth	to	14	puppies	!
	LOCATION	ANIMAL	BODY	PATIENT		ANIMAL	

the expression *chocolate lab* is marked as a fundamental unit of meaning—a **multiword expression**—whose supersense label indicates that it is an animal, not a confectionary research facility.

Designing this style of analysis—and then automating it—broke ground in the following respects:

- Devising a *comprehensive* scheme for annotating multiword expressions (Schneider et al., 2014b); previous work had addressed only certain kinds (for a review, see Baldwin and Kim, 2010).
- Efficiently capturing discontinuous lexical expressions (e.g., *sniff something out*) in a discriminative sequence chunking model: we discovered a simple solution based on adapting BIO-style tags without sacrificing linear-time exact search (Schneider et al., 2014a).
- Enumerating precise linguistic criteria for each of the noun and verb supersense categories, demonstrating they are suitable for direct annotation and are not specific to English (Schneider et al., 2012; Schneider and Smith, 2015).
- Developing a new, linguistically-based, hierarchical supersense inventory for **prepositions** (Schneider et al., 2015) and comprehensively annotating them in a corpus (Schneider et al., submitted).
- Integrating supersenses and comprehensive multiword expressions in a joint sequence model (Schneider and Smith, 2015).

Impact. My work has produced a 55,000 word corpus annotated with lexical semantic analyses, which is publicly available on the web. This corpus has prompted an MSc thesis project on detecting inconsistencies in semantic annotation, and another on integrating lexical semantics into machine translation. Moreover, I am coordinating (along with Dirk Hovy, Anders Johannsen, and Marine Carpuat) a shared task competition that builds on our problem formulation and dataset: systems must predict lexical semantic segments and supersenses for each sentence. The task is expected to attract a number of system submissions, which will be compared on multiple genres of text.

Ongoing and future work. Multiword expressions and prepositions are both phenomena blurring the line between lexicon and grammar, so analyzing them opens up interesting possibilities for study with regard to language learning, translation, and linguistic theory. For example, I am planning to annotate prepositions in child language utterances to test hypotheses about syntactic vs. semantic acquisition patterns. Related to prepositions, I am interested in the semantic and discourse functions of closed-class grammatical phenomena, such as definiteness marking (Bhatia et al., 2014). Finally, supersenses can be extended to cover additional categories of content words, and we have started to map them out for adjectives (Tsvetkov et al., 2014).

3 Relational Semantics

Analyzing sentences with richer semantic abstractions. I have worked with two representations of events and other relations that express, loosely speaking, “who did what to whom, under what circumstances”. These describe how the words within a sentence relate to one another to form complex meanings, which must go well beyond syntax (any meaning representation which makes *having a baby* look more similar to *having a cookie* than to *giving birth* is clearly missing something!). One set of contributions were new machine learning models for predicting predicate-argument structures based on the **FrameNet** lexical resource (Fillmore and Baker, 2009): these models take better advantage of FrameNet-annotated training sentences (see figure 1), and also provide ways to improve predictions using auxiliary data resources (Das et al., 2010, 2014; Kshirsagar et al., 2015). Second, I helped design the **Abstract Meaning**

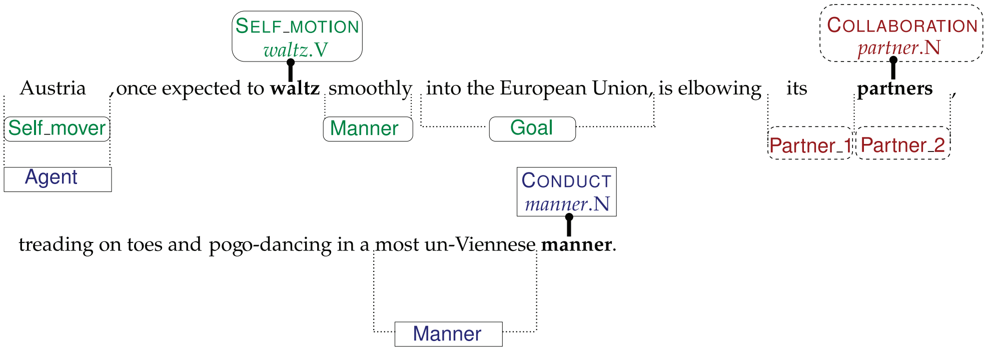
SELF_MOTION
waltz.V
COLLABORATION
partner.N
Austria , once expected to waltz smoothly into the European Union, is elbowing its partners ,
Self_mover
Agent
Manner
Goal
Partner_1
Partner_2
CONDUCT
manner.N
treading on toes and pogo-dancing in a most un-Viennese manner.
Manner

Figure 1: A sentence with its frame-semantic parse (Das et al., 2014). The FrameNet lexicon lists frames (conceptual scenarios) and their roles (argument slots); for example, the COLLABORATION frame calls for two individuals, `Partner_1` and `Partner_2`. The SEMAFOR system labels substrings of an input sentence with these frame names and arguments.

Representation (AMR) formalism for large-scale domain-general annotation of English sentence semantics (Banarescu et al., 2013). AMR expresses a sentence's meaning in a graph, where nodes represent concepts (events, entities, attributes) and edges represent relations (`part-of`, `agent-of`, `location-of`, etc.). I see FrameNet and AMR as complementary, the former providing deeper lexical semantics and the latter expressing a greater number of compositional relations.

Impact. Our frame-semantic parser, **SEMAFOR**, has been downloaded over a thousand times since its initial release. In addition to further studies of semantic role labeling, other researchers have used SEMAFOR for predicting changes in stock prices (Xie et al., 2013) and slot-filling for spoken dialogue systems (Chen et al., 2013), among other things. **AMR** has stimulated a flurry of research in graph automata (Jones et al., 2012; Chiang et al., 2013; Koller, 2015), graph-semantic parsing (Flanigan et al., 2014; Werling et al., 2015; Pust et al., 2015; Artzi et al., 2015, *inter alia*) and summarization (Liu et al., 2015), and cross-linguistic semantic variation (Xue et al., 2014; Vanderwende et al., 2015). Tens of thousands of English sentences are being annotated with AMRs to facilitate large-scale statistical parsing, generation, and applications including machine translation.

4 Upgrading NLP for the Social Web

Analyzing text from Twitter, Wikipedia, and online reviews. In the realm of syntax, I have worked on building datasets and analyzers for **English Twitter text**: part-of-speech tagging (Gimpel et al., 2011; Owoputi et al., 2013) and unlabeled dependency parsing (Kong et al., 2014). The style of language on Twitter differs sharply from expository writing seen in news or even Wikipedia articles: messages are short, opinionated, and often contain slang/dialectal language and unorthodox spellings. These properties tend to throw off conventional NLP tools trained on conventional language, so we built a Twitter dependency treebank in order to train a Twitter-friendly parser. Moreover, forms of linguistic annotation designed for traditional genres need to be adapted or replaced for new genres, as different syntactic and discourse phenomena will be present (Schneider, 2015). Our corpus leverages a novel annotation scheme for economical descriptions of syntactic structure in such genres (Schneider et al., 2013b).

Much of my work on lexical semantics has likewise targeted online language, including datasets from **Arabic language Wikipedia** (Mohit et al., 2012; Schneider et al., 2012, 2013a) and **online reviews**

(Schneider et al., 2014a,b; Schneider and Smith, 2015, Schneider et al., submitted).

Impact. There is considerable demand for NLP for online genres, with application to such tasks as sentiment analysis/opinion mining, event detection/prediction, and disaster response. Our tools have been widely used—e.g., the POS tagger has been downloaded over 6,000 times. Our dependency annotation framework has facilitated datasets and parsers for low-resource languages such as Kinyarwanda (Mielens et al., 2015). The shared task mentioned above (§2) will target domains including Twitter and online reviews.

5 Outlook

Looking forward, I will continue to pursue the twin goals of *collecting* and *exploiting* linguistically analyzed data in ways that are more accurate, more efficient, and more robust. Some directions of interest:

Scaling up structured event lexicons. The meaning representations discussed in §3 depend on event lexicons such as FrameNet to map lexical predicates to abstract, role-defining senses/scenarios. These lexicons are traditionally constructed by hand, which can require substantial expertise and effort—both of which grow with the vocabulary coverage and the extent of abstraction. To build event lexicons at scale without sacrificing their expressiveness will require new sources of knowledge to complement or support the work of language and domain experts. Some relevant information may come from naïve human judgments (such as through crowdsourcing). I am currently investigating distributional methods for discovering predicate relationships from data. Optimizing the balance of different sources of information, with different costs and degrees of reliability, is a foundational challenge of data science as applied to AI.

Semi-supervised and multi-task learning to exploit heterogeneous data for semantic analysis. The frame-semantic parsing work described above has incorporated these machine learning techniques to an extent, but there are further opportunities for jointly leveraging several different kinds of relational semantic data, including PropBank, AMR, and FrameNet corpora. To improve lexical semantic analysis, I would like to find ways to exploit partially annotated resources such as SemCor and OntoNotes, parallel data, and unlabeled data in combination with my annotated training data.

Models that more tightly couple morphology, syntax, and semantics. My work on frame-semantic parsing discussed above models a sentence's semantic structure conditional on features derived in part from a syntactic parse. I am interested in whether broad-coverage syntax and relational semantics are better modeled by making their relationship more explicit, as in CCG (Steedman, 2000; Artzi et al., 2015) and some other grammar formalisms. I believe the theoretical framework of Construction Grammar (Hoffmann and Trousdale, 2013) offers valuable insights, of which NLP models have barely scratched the surface (Schneider and Tsarfaty, 2013). Additionally, I want to build models that take morphology seriously, especially for languages other than English. Having worked (from a cognitive and constructional perspective) on the morphological, morphosyntactic, and semantic dimensions of Hebrew verbs (Schneider, 2010), I think there is utility in modeling phenomena that cut across these levels of structure.

Corpus-based tools for studying human language acquisition and assisting language learners. Language use by nonnative speakers, as opposed to native speakers, is an important source of evidence about native speaker knowledge. I have worked on detecting the native languages of ESL speakers based on cues in their English writing (Tsvetkov et al., 2013), which can be done with high accuracy, suggesting that technologies to encourage native-like writing are within reach. My dissertation's modeling of lexical semantics should prove useful here—prepositions and multiword idioms are notoriously difficult for second language learners. Learner corpora, as a kind of "nonstandard" language domain, present interesting challenges for linguistic structure NLP as well.

References

Y. Artzi, K. Lee, and L. Zettlemoyer. Broad-coverage CCG semantic parsing with AMR. In *Proc. of EMNLP*, 2015.

T. Baldwin and S. N. Kim. Multiword expressions. In N. Indurkhya and F. J. Damerau, editors, *Handbook of Natural Language Processing, Second Edition*. CRC Press, Taylor and Francis Group, Boca Raton, FL, 2010.

L. Banarescu, C. Bonial, S. Cai, M. Georgescu, K. Griffitt, U. Hermjakob, K. Knight, P. Koehn, M. Palmer, and N. Schneider. Abstract Meaning Representation for sembanking. In *Proc. of LAW*, 2013.

A. Bhatia, C.-C. Lin, N. Schneider, Y. Tsvetkov, F. Talib Al-Raisi, L. Roostapour, J. Bender, A. Kumar, L. Levin, M. Simons, and C. Dyer. Automatic classification of communicative functions of definiteness. In *Proc. of COLING*, 2014.

Y.-N. Chen, W. Y. Wang, and A. I. Rudnicky. Unsupervised induction and filling of semantic slots for spoken dialogue systems using frame-semantic parsing. In *Proc. of the IEEE ASRU 2013*, 2013.

D. Chiang, J. Andreas, D. Bauer, K. M. Hermann, B. Jones, and K. Knight. Parsing Graphs with Hyperedge Replacement Grammars. In *Proc. of ACL*, 2013.

M. Ciaramita and Y. Altun. Broad-coverage sense disambiguation and information extraction with a supersense sequence tagger. In *Proc. of EMNLP*, 2006.

D. Das, N. Schneider, D. Chen, and N. A. Smith. Probabilistic frame-semantic parsing. In *Proc. of NAACL-HLT*, 2010.

D. Das, D. Chen, A. F. T. Martins, N. Schneider, and N. A. Smith. Frame-semantic parsing. *Computational Linguistics*, 40(1), 2014.

C. J. Fillmore and C. Baker. A frames approach to semantic analysis. In B. Heine and H. Narrog, editors, *The Oxford Handbook of Linguistic Analysis*, pages 791–816. Oxford University Press, Oxford, UK, 2009.

J. Flanigan, S. Thomson, J. Carbonell, C. Dyer, and N. A. Smith. A discriminative graph-based parser for the Abstract Meaning Representation. In *Proc. of ACL*, 2014.

K. Gimpel, N. Schneider, B. O'Connor, D. Das, D. Mills, J. Eisenstein, M. Heilman, D. Yogatama, J. Flanigan, and N. A. Smith. Part-of-speech tagging for Twitter: annotation, features, and experiments. In *Proc. of ACL-HLT*, 2011.

T. Hoffmann and G. Trousdale, editors. *The Oxford Handbook of Construction Grammar*. Oxford University Press, Oxford, UK, 2013.

B. Jones, J. Andreas, D. Bauer, K. M. Hermann, and K. Knight. Semantics-based machine translation with hyperedge replacement grammars. In *Proc. of COLING 2012*, 2012.

A. Koller. Semantic construction with graph grammars. In *Proc. of the 11th International Conference on Computational Semantics*, 2015.

L. Kong, N. Schneider, S. Swayamdipta, A. Bhatia, C. Dyer, and N. A. Smith. A dependency parser for tweets. In *Proc. of EMNLP*, 2014.

M. Kshirsagar, S. Thomson, N. Schneider, J. Carbonell, N. A. Smith, and C. Dyer. Frame-semantic role labeling with heterogeneous annotations. In *Proc. of ACL-IJCNLP*, 2015.

F. Liu, J. Flanigan, S. Thomson, N. Sadeh, and N. A. Smith. Toward abstractive summarization using semantic representations. In *Proc. of NAACL-HLT*, 2015.

J. Mielens, L. Sun, and J. Baldridge. Parse imputation for dependency annotations. In *Proc. of ACL-IJCNLP*, 2015.

B. Mohit, N. Schneider, R. Bhowmick, K. Oflazer, and N. A. Smith. Recall-oriented learning of named entities in Arabic Wikipedia. In *Proc. of EACL*, 2012.

O. Owoputi, B. O'Connor, C. Dyer, K. Gimpel, N. Schneider, and N. A. Smith. Improved part-of-speech tagging for online conversational text with word clusters. In *Proc. of NAACL-HLT*, 2013.

M. Pust, U. Hermjakob, K. Knight, D. Marcu, and J. May. Parsing English into Abstract Meaning Representation using syntax-based machine translation. In *Proc. of EMNLP*, 2015.

N. Schneider. Computational cognitive morphosemantics: modeling morphological compositionality in Hebrew verbs with Embodied Construction Grammar. In *Proc. of BLS*, 2010.

N. Schneider. What I've learned about annotating informal text (and why you shouldn't take my word for it). In *Proc. of LAW*, 2015.

N. Schneider. *Lexical Semantic Analysis in Natural Language Text.* Ph.D. dissertation, Carnegie Mellon University, Pittsburgh, Pennsylvania, 2014.

N. Schneider and N. A. Smith. A corpus and model integrating multiword expressions and supersenses. In *Proc. of NAACL-HLT*, 2015.

N. Schneider and R. Tsarfaty. Design Patterns in Fluid Construction Grammar, Luc Steels (editor). *Computational Linguistics*, 39(2):447–453, 2013.

N. Schneider, B. Mohit, K. Oflazer, and N. A. Smith. Coarse lexical semantic annotation with supersenses: an Arabic case study. In *Proc. of ACL*, 2012.

N. Schneider, B. Mohit, C. Dyer, K. Oflazer, and N. A. Smith. Supersense tagging for Arabic: the MT-in-the-middle attack. In *Proc. of NAACL-HLT*, 2013a.

N. Schneider, B. O'Connor, N. Saphra, D. Bamman, M. Faruqui, N. A. Smith, C. Dyer, and J. Baldridge. A framework for (under)specifying dependency syntax without overloading annotators. In *Proc. of LAW*, 2013b.

N. Schneider, E. Danchik, C. Dyer, and N. A. Smith. Discriminative lexical semantic segmentation with gaps: running the MWE gamut. *Trans. ACL*, 2014a.

N. Schneider, S. Onuffer, N. Kazour, E. Danchik, M. T. Mordowanec, H. Conrad, and N. A. Smith. Comprehensive annotation of multiword expressions in a social web corpus. In *Proc. of LREC*, 2014b.

N. Schneider, V. Srikumar, J. D. Hwang, and M. Palmer. A hierarchy with, of, and for preposition supersenses. In *Proc. of LAW*, 2015.

N. Schneider, V. Srikumar, J. D. Hwang, M. Green, T. O'Gorman, K. Conger, and M. Palmer. Hierarchical preposition supersense tagging. Submitted.

M. Steedman. *The Syntatic Process.* MIT Press, Cambridge, MA, 2000.

Y. Tsvetkov, N. Twitto, N. Schneider, N. Ordan, M. Faruqui, V. Chahuneau, S. Wintner, and C. Dyer. Identifying the L1 of non-native writers: the CMU-Haifa system. In *Proc. of BEA*, 2013.

Y. Tsvetkov, N. Schneider, D. Hovy, A. Bhatia, M. Faruqui, and C. Dyer. Augmenting English adjective senses with supersenses. In *Proc. of LREC*, 2014.

L. Vanderwende, A. Menezes, and C. Quirk. An AMR parser for English, French, German, Spanish and Japanese and a new AMR-annotated corpus. In *Proc. of NAACL-HLT: Demonstrations*, 2015.

K. Werling, G. Angeli, and C. D. Manning. Robust subgraph generation improves Abstract Meaning Representation parsing. In *Proc. of ACL-IJCNLP*, 2015.

B. Xie, R. J. Passonneau, L. Wu, and G. G. Creamer. Semantic frames to predict stock price movement. In *Proc. of ACL*, 2013.

N. Xue, O. Bojar, J. Hajič, M. Palmer, Z. Urešová, and X. Zhang. Not an interlingua, but close: comparison of English AMRs to Chinese and Czech. In *Proc. of LREC*, 2014.