Lexical Semantic Analysis in Natural Language Text

EXTENDED ABSTRACT

Nathan Schneider

1 Introduction

The central challenge in computational lexical semantics for text corpora is to develop and apply abstractions that characterize word meanings beyond what can be derived superficially from the orthography. Such abstractions can be found in type-level human-curated lexical resources such as WordNet (Fellbaum, 1998), but such intricate resources are expensive to build and difficult to annotate with at the token level, hindering their applicability beyond a narrow selection of languages and domains. For empirical study and real-world NLP applications in a wide range of text corpora, a more portable and scalable—yet still linguistically-grounded—way to represent lexical meanings is needed.

This thesis formalizes a scheme for robust description and modeling of lexical expressions in text. This includes an annotation scheme that makes it practical for humans to rapidly specify lexical semantic units and classes at the token level, achieving broad coverage without any dependency on a lexicon or syntactic parse (which would hinder coverage in low-resource languages and domains). For each sentence, the scheme specifies (i) **multiword expressions**, resulting in a *lexical semantic segmentation*, and (ii) **supersense** classes for nouns, verbs, and prepositions. Each aspect of the scheme draws on previous work but is novel in important respects. The components are integrated in a formal representation that facilitates supervised learning and prediction with statistical sequence models. We apply the scheme in an empirical case study to informal English text (online reviews). This case study validates the descriptive approach and the manual annotation methodology, quantifies the effects of statistical modeling deci-

sions, and provides a corpus resource and automatic tool for the benefit of future studies.

Following an introduction and general literature survey, the thesis devotes three chapters to describing the scheme and annotation process, and another two chapters to statistical tagging experiments. The section numbers and titles below correspond to chapters of the thesis. Citations for publications with content from the chapter are shown alongside the title.

2 General Background: Computational Lexical Semantics

This work exists against the backdrop of two dominant paradigms for computational lexical semantics. The first is **word sense disambiguation** (WSD), in which it is typically assumed that a lexicon (such as WordNet) defines sense refinements for word types. The disambiguation task is to select, using contextual information, the appropriate sense for each token of that type.

The other paradigm is **named entity recognition** (NER), the detection and coarse classification of proper name mentions in context. Canonical NER systems are expected to generalize beyond particular name types seen in training data. Often, a tagger is trained to jointly identify mention boundaries and choose a general-purpose class label (such as PERSON, ORGANIZATION, or LOCATION) for each mention.

Each of these approaches has disadvantages. Traditional WSD is fine-grained and lexicon-dependent, making annotation of new datasets difficult and limiting portability to new domains and languages. On the other hand, NER exploits coarse-grained and general-purpose classes, which makes corpus annotation cost-effective—but its coverage (by definition) is limited to proper names.

This work advocates an approach to lexical semantic representation that is coarse-grained, leveraging the advantages of NER, but applies to most content words (not just names), approximating the reach of WSD in a way that does not depend on lexicon coverage.

We describe the components of the representation in turn.

3 Multiword Expressions (Schneider et al., 2014b)

Multiword expressions (MWEs) are both *numerous*, occurring frequently in text, and *diverse*—they are not restricted to particular syntactic construc-

tions or semantic domains (Baldwin and Kim, 2010). This thesis introduces a comprehensive and broad-coverage framework for representing diverse MWEs in corpora, without requiring a lexicon.

Our approach to MWEs in context is *comprehensive*, meaning that it is not restricted to a particular lexical or even syntactic inventory of candidates. Included are the full spectrum of MWE classes—ranging from the most fixed (proper names, nominal compounds, connectives like *as well as*, idioms like *by and large*) to the most flexible (especially verb phrase expressions subject to internal modification or other syntactic processes affecting word order and/or contiguity). For example, the expression whose citation form is *pay attention to* could be instantiated as ***paid*** *no* ***attention to*** or ***attention*** *was* ***paid to***, both of which contain **gaps** between the lexicalized parts of the expression. Further, the object of the preposition is not part of the MWE, so the MWE is not a complete constituent by a standard syntactic analysis.

The approach taken here is to bypass the difficult issue of syntactic representation altogether: the (shallow) MWE representation simply assigns tokens to groups, where each group reflects the lexicalized part of an MWE. Tokens within a group are not required to be contiguous.

The following example is a fragment of an online review whose tokens have been grouped into lexical expressions, each identified with an index:

(1) I googled restaurants in the area and **Fuji Sushi** *came up* and
1 2 3 4 5 6 7 8 8 9 9 10
reviews were great so I **made** a *carry out* **order**
11 12 13 14 15 16 17 18 18 17

The four multiword expressions are *Fuji Sushi*, *came up*, *made… order*, and *carry out*. Notably, *carry out* occurs inside the gap of *made… order* (along with the indefinite article, which is not part of either MWE.)

Because it can be difficult to draw a sharp line between MWEs and productive combinations, we further represent two degrees of MWE-hood. In our full scheme, there are **weak** groups for statistically idiomatic collocations, such as *highly recommended*, and **strong** groups for all MWEs involving an element of noncompositionality. A strong group may include one or more weak groups, but otherwise there is no nesting of groups.

To facilitate automatic sequence tagging, the group annotations are mapped to an encoding similar to the traditional BIO scheme for chunking (Ramshaw and Marcus, 1995): namely, 8 tags—`O`, `o`, `B`, `b`, `Ī`, `ī`, `Ĩ`, `ĩ`—allow for the distinctions of tokens:

- positioned in the gap of some MWE (lowercase tags) vs. not (uppercase

tags), and

- not belonging to an MWE (`O/o`), beginning an MWE (`B/b`), continuing a strong MWE (`Ī/ī`), or continuing a weak MWE (`Ĩ/ĩ`).

This encoding allows for MWEs with multiple gaps (e.g., ***putting*** *me* ***at*** *my* ***ease***). It prohibits any MWE with a gap from occurring in the gap of another MWE, and also excludes weak MWEs consisting of a gappy strong MWE and one or more tokens inside the gap. That these constraints are linguistically reasonable is empirically supported by the annotated corpus (counterexamples exist but are extremely rare). Importantly for exactness of dynamic programming inference, a well-formed segmentation of the sentence can be enforced with bigram constraints on tag transitions.

Here is an example of the 8-tag scheme:

(2) The white pages allowed me to get in touch with parents of my
B Ī O O O B Ĩ Ī Ĩ O O O
high school friends so that I could track people down one by one
B Ī O O O O O B o Ī B Ī Ī

This example is annotated with 4 contiguous strong MWEs (*white pages*, *in touch*, *high school*, *one by one*), one gappy strong MWE (*track…down*), and a contiguous weak MWE (*get in touch with*)—note that a subset of the weak MWE's tokens also form a strong MWE.

Annotation. A corpus of 723 online reviews from the English Web Treebank (Bies et al., 2012) has been annotated in this framework. The text in this corpus is written in an informal style and colloquial idioms are frequent. The comprehensive annotations cover 3,800 sentences (55k words); 3,024 strong and 459 weak MWEs are annotated. Each sentence was independently annotated by at least two annotators, who then negotiated a consensus for any disagreements. All annotators hold bachelor's degrees in linguistics. Inter-annotator agreement is quantified and deemed acceptable. Because the corpus is from a treebank, the shallow MWE annotations can be aligned post hoc to syntactic parses for future compositional models.

4 Noun and Verb Supersenses

(Schneider et al., 2012; Schneider and Smith, 2015)

The second part of our lexical semantic representation is to assign semantic labels to the units. Applying fine-grained word senses (as in classical word sense disambiguation) depends on the coverage of a large sense inventory,

Noun			Verb		
GROUP	1469	*place*	STATIVE	2922	*is*
PERSON	1202	*people*	COGNITION	1093	*know*
ARTIFACT	971	*car*	COMMUNIC.*	974	*recommend*
COGNITION	771	*way*	SOCIAL	944	*use*
FOOD	766	*food*	MOTION	602	*go*
ACT	700	*service*	POSSESSION	309	*pay*
LOCATION	638	*area*	CHANGE	274	*fix*
TIME	530	*day*	EMOTION	249	*love*
EVENT	431	*experience*	PERCEPTION	143	*see*
COMMUNIC.*	417	*review*	CONSUMPTION	93	*have*
POSSESSION	339	*price*	BODY	82	*get…done*
ATTRIBUTE	205	*quality*	CREATION	64	*cook*
QUANTITY	102	*amount*	CONTACT	46	*put*
ANIMAL	88	*dog*	COMPETITION	11	*win*
BODY	87	*hair*	WEATHER	0	—
STATE	56	*pain*	**all 15 VSSTs**	**7806**	
NATURAL OBJ.	54	*flower*			
RELATION	35	*portion*	**N/A** (see text)		
SUBSTANCE	34	*oil*	`a	1191	*have*
FEELING	34	*discomfort*	`	821	*anyone*
PROCESS	28	*process*	`j	54	*fried*
MOTIVE	25	*reason*			
PHENOMENON	23	*result*	*COMMUNIC. *is short for* COMMUNICATION		
SHAPE	6	*square*			
PLANT	5	*tree*			
OTHER	2	*stuff*			
all 26 NSSTs	**9018**				

Table 1: Summary of noun and verb supersense categories. Each entry shows the label along with the count and most frequent lexical item in the **STREUSLE** corpus.

which we want to avoid so as to ease adaptation to new domains and languages. Moreover, fine-grained sense annotation is slow and difficult for humans, and therefore costly if high-quality annotations are desired.

Therefore, we instead use unlexicalized coarse-grained labels called **supersenses**. For nouns and verbs, we take the inventory of categories specified by WordNet (table 1) and repurpose it for direct annotation. This involved writing definitions for the categories and providing guidance to help annotators apply them consistently. For nouns, our conventions have been applied both to Arabic Wikipedia articles and English web reviews, demonstrating the robustness of the supersense classes. (We have applied verb categories to the English data but expect that they would generalize similarly.)

The white pages allowed me to get in touch with
$_B$N:COMMUNICATION Ĩ $_O$V:COGNITION O O $_B$V:SOCIAL Ĩ Ĩ Ĩ
parents of my high school friends so that I could
$_O$N:PERSON O O $_B$N:GROUP Ĩ $_O$N:PERSON O O O O
track people down one by one
$_B$V:SOCIAL $_O$N:PERSON Ĩ B Ĩ Ĩ

Figure 1: Supersense tagging on top of the lexical semantic segmentation of (2). Note that the supersense label is only attached to the first tag of the expression.

For English web reviews, we have augmented the aforementioned MWE dataset, adding gold supersense labels for nouns and verbs. Because strong MWEs function as a lexical semantic unit, they receive no more than one supersense: e.g., *pay attention* would be annotated holistically as V:COGNITION. In the tag representation of MWEs, the supersense is simply appended to the first tag in the lexical expression. The supersense-enriched version of (2) is thus encoded as shown in figure 1.

Table 1 shows counts of each supersense in our dataset, which has been released at http://www.ark.cs.cmu.edu/LexSem/ under the name **STREUSLE**.[1]

5 Preposition Supersenses (Schneider et al., 2015)

In principle, the assignment of lexical semantic classes need not be limited to nouns and verbs. This thesis extends the inventory of supersenses (unlexicalized categories) to include prepositions. A handful of preposition types (*of*, *to*, *in*, etc.) are extremely frequent and important as arbiters of semantic relations. They are also extremely polysemous, and their polysemy patterns are language-specific—so being able to disambiguate prepositions in context is necessary for translation. For example, here is just a sample of the functions of (prepositional or infinitival) *to* in English:

(3) a. My cake is **to** die for. *(nonfinite verb idiom)*
b. If you want I can treat you **to** some. *(prepositional verb idiom)*
c. How about this: you go **to** the store *(locative goal)*
d. **to** buy ingredients. *(nonfinite purpose)*
e. That part is up **to** you. *(responsibility)*
f. Then if you give the recipe **to** me *(recipient)*
g. I'm happy **to** make the batter *(nonfinite adjectival complement)*

[1] Supersense-Tagged Repository of English with a Unified Semantics for Lexical Expressions

h. and put it in the oven for 30 **to** 40 minutes *(range limit)*
i. so you will arrive **to** the sweet smell of chocolate. *(background event)*
j. That sounds good **to** me. *(affective/experiencer)*
k. I hope it lives up **to** your expectations. *(prepositional verb idiom)*
l. That's all there is **to** it. *(phrasal idiom)*

To date, computational accounts of English preposition semantics have not been ideal for rapid and broad-coverage annotation. The Preposition Project (TPP; Litkowski and Hargraves, 2005) documents fine-grained lexicographic senses, but annotation with these is slow as with fine-grained WordNet senses. Srikumar and Roth (2013) proposed an inventory of unlexicalized preposition classes based on clustering the TPP senses—we took this inventory as a starting point, but found that many of the categories had unclear boundaries or seemed to conflate multiple phenomena in a way that confused annotators. A new inventory, structured as a multiple inheritance hierarchy modeled after VerbNet's (Bonial et al., 2011), is proposed in this thesis. Unlike some of the previous work on prepositions, our approach takes into account (a) multiword expressions functioning as prepositions (e.g., *out of*, *due to*)—these receive a preposition supersense holistically; and (b) prepositions/particles that are part of a larger multiword expression (e.g., *make up* 'invent', *be to die for*)—these do not receive a preposition supersense because the MWE does not function as a preposition.

Figure 2 illustrates the portion of the hierarchy devoted to temporal prepositions. Many of the category labels are familiar from VerbNet's thematic roles, though some are more specific. Note that most categories are associated with a small number of preposition types. In addition to the hierarchy, this work provides a resource documenting known type–supersense mappings, with example sentences for each mapping. This enables annotators to see a filtered list of options for the preposition they are annotating. Though the preposition supersense inventory is more complex than the noun and verb inventories, in pilot annotation studies we have found that it is feasible with the resource. Full annotation of the prepositions in the English web reviews corpus is ongoing.

6 Multiword Expression Identification

(Schneider et al., 2014a)

The shallow MWE annotations described above can, with the 8-tag encoding, be used to train and evaluate a statistical sequence tagger, similar to other shallow MWE identification systems (e.g., Constant et al., 2012). The

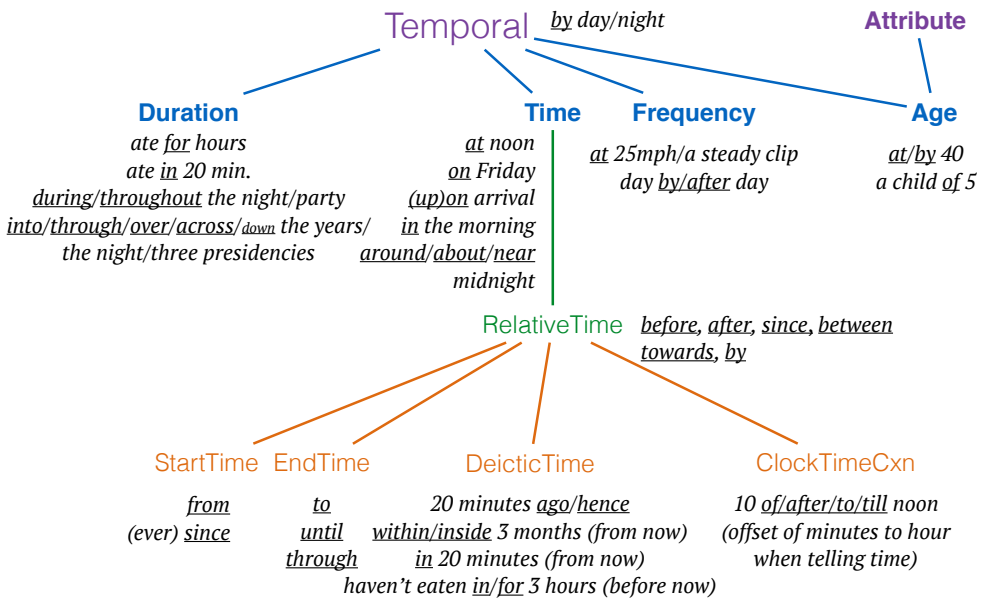

Figure 2: The temporal subhierarchy, with example preposition usages associated with each supersense.

POS pattern	#	examples (lowercased lemmas)
NOUN NOUN	53	*customer service, oil change*
VERB PREP	36	*work with, deal with, yell at*
PROPN PROPN	29	*eagle transmission, comfort zone*
ADJ NOUN	21	*major award, top notch*
VERB PART	20	*move out, end up, pick up, pass up*
VERB ADV	17	*come back, come in, come by*
PREP NOUN	12	*on time, in fact, in cash, for instance*
VERB NOUN	10	*take care, make money, give crap*
VERB PRON	10	*thank you, get it*
PREP PREP	8	*out of, due to, out ta, in between*
ADV ADV	6	*no matter, up front, at all, early on*
DET NOUN	6	*a lot, a little, a bit, a deal*
VERB DET NOUN	6	*answer the phone, take a chance*
NOUN PREP	5	*kind of, care for, tip on, answer to*

Table 2: Top predicted POS patterns and counts.

strong vs. weak distinction and the way in which gaps are allowed are novel. We adapt the feature representation of Constant et al. (2012), incorporating several MWE lexicons and training a discriminative first-order Markov model with the structured perceptron (Collins, 2002).

Evaluation measure. We quantify inter-annotator agreement and system comparisons with a new evaluation measure for automatic shallow MWE analyses. The main idea is that partial credit is given for partial overlap between a gold MWE instance and a predicted MWE instance by computing precision and recall not over the full MWE, but over *links* between consecutive tokens belonging to each MWE. Weak MWEs are treated as intermediate between strong and non-MWEs, so minor disagreements about the strength level are not punished too harshly.

Experimental results. Experiments on held-out data show that the statistical model is vastly superior to a baseline involving heuristic matching against MWE lexicons. However, utilizing those lexicons in a soft way, through features, is beneficial. The best result (without gold POS tags) is 64% precision, 56% recall, and 59% F_1 on the test set. Table 2 shows a sample of the system's output.

7 Full Supersense Tagging (Schneider and Smith, 2015)

Having built a corpus annotated for comprehensive multiword expressions as well as noun and verb supersenses, it is possible to train a sequence tagger that jointly predicts both representational components. We present a model that is similar to the supersense tagger of Ciaramita and Altun (2006), but trained on our corpus and also identifying a broad range of multiword expressions.

A joint model is achieved by representing both MWE positional information and supersense labels in the same tag space. With $|\mathcal{N}| = 26$ noun supersense classes and $|\mathcal{V}| = 16$ verb classes, there are in principle

$$\underbrace{|\{\texttt{O o B b \tilde{I} \tilde{i}}\}|}_{6} \times \underbrace{(1 + |\mathcal{N}| + |\mathcal{V}|)}_{43} + \underbrace{|\{\bar{\texttt{I}}\ \bar{\texttt{i}}\}|}_{2} = 260$$

possible tags encoding chunk and class information, allowing for chunks with no class because they are neither nominal nor verbal expressions. In practice, though, many of these combinations are nonexistent in our data; for experiments we only consider tags occurring in the training data, yielding $|\mathcal{Y}| = 146$. Exact decoding with the Viterbi algorithm is linear in the length of the sentence and quadratic in the number of tags; this produces runtimes well within the range of practicality for our dataset, so approximate inference techniques are not needed. Our full model attains 71% supersense labeling F_1 with almost no impact on MWE identification performance.

8 Conclusion

This thesis has provided a framework for describing the lexical units and semantic classes within text sentences, manually and automatically, with broad coverage. Because the general framework does not depend on any pre-existing lexical resource, it is expected to be suitable for a wide range of text domains and languages. The thesis has motivated and detailed innovations in the *representation* of lexical semantics, a practical approach to human *annotation* of corpora, and statistical techniques for the *automation* of the analysis using said corpora. The primary case study concerning sentences from English web reviews allowed for each of these steps to be understood and documented qualitatively and quantitatively. It has also produced an annotated corpus resource and analysis software, both of which will be released to facilitate further linguistic investigation, computational modeling, and application to other tasks (such as semantic parsing and machine translation).

Bibliography

Timothy Baldwin and Su Nam Kim. Multiword expressions. In Nitin Indurkhya and Fred J. Damerau, editors, *Handbook of Natural Language Processing, Second Edition*, pages 267–292. CRC Press, Taylor and Francis Group, Boca Raton, Florida, USA, 2010.

Ann Bies, Justin Mott, Colin Warner, and Seth Kulick. English Web Treebank. Technical Report LDC2012T13, Linguistic Data Consortium, Philadelphia, PA, 2012. URL http://www.ldc.upenn.edu/Catalog/catalogEntry.jsp?catalogId=LDC2012T13.

Claire Bonial, William Corvey, Martha Palmer, Volha V. Petukhova, and Harry Bunt. A hierarchical unification of LIRICS and VerbNet semantic roles. In *Fifth IEEE International Conference on Semantic Computing*, pages 483–489, Palo Alto, CA, USA, September 2011.

Massimiliano Ciaramita and Yasemin Altun. Broad-coverage sense disambiguation and information extraction with a supersense sequence tagger. In *Proc. of EMNLP*, pages 594–602, Sydney, Australia, July 2006.

Michael Collins. Discriminative training methods for Hidden Markov Models: theory and experiments with perceptron algorithms. In *Proc. of EMNLP*, pages 1–8, Philadelphia, Pennsylvania, USA, July 2002.

Matthieu Constant, Anthony Sigogne, and Patrick Watrin. Discriminative strategies to integrate multiword expression recognition and parsing. In *Proc. of ACL*, pages 204–212, Jeju Island, Korea, July 2012.

Christiane Fellbaum, editor. *WordNet: an electronic lexical database*. MIT Press, Cambridge, Massachusetts, USA, 1998.

Ken Litkowski and Orin Hargraves. The Preposition Project. In *Proc. of the Second ACL-SIGSEM Workshop on the Linguistic Dimensions of Prepositions and their Use in Computational Linguistics Formalisms and Applications*, pages 171–179, Colchester, Essex, UK, April 2005.

Lance A. Ramshaw and Mitchell P. Marcus. Text chunking using transformation-based learning. In *Proc. of the Third ACL Workshop on Very Large Corpora*, pages 82–94, Cambridge, Massachusetts, USA, June 1995.

Nathan Schneider and Noah A. Smith. A corpus and model integrating multiword expressions and supersenses. In *Proc. of NAACL-HLT*, Denver, Colorado, USA, June 2015.

Nathan Schneider, Behrang Mohit, Kemal Oflazer, and Noah A. Smith. Coarse lexical semantic annotation with supersenses: an Arabic case study. In *Proc. of ACL*, pages 253–258, Jeju Island, Korea, July 2012.

Nathan Schneider, Emily Danchik, Chris Dyer, and Noah A. Smith. Discriminative lexical semantic segmentation with gaps: running the MWE gamut. *Transactions of the Association for Computational Linguistics*, 2:193–206, April 2014a.

Nathan Schneider, Spencer Onuffer, Nora Kazour, Emily Danchik, Michael T. Mordowanec, Henrietta Conrad, and Noah A. Smith. Comprehensive annotation of multiword expressions in a social web corpus. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Hrafn Loftsson, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proc. of LREC*, pages 455–461, Reykjavík, Iceland, May 2014b.

Nathan Schneider, Vivek Srikumar, Jena D. Hwang, and Martha Palmer. A hierarchy with, of, and for preposition supersenses. In *Proceedings of the 9th Linguistic Annotation Workshop*, Denver, Colorado, USA, June 2015.

Vivek Srikumar and Dan Roth. An inventory of preposition relations. Technical Report arXiv:1305.5785, May 2013. URL http://arxiv.org/abs/1305.5785.