A Corpus and Model Integrating Multiword Expressions and Supersenses

Nathan Schneider
School of Informatics
University of Edinburgh
Edinburgh, Scotland, UK
nschneid@inf.ed.ac.uk

Noah A. Smith
School of Computer Science
Carnegie Mellon University
Pittsburgh, Pennsylvania, USA
nasmith@cs.cmu.edu

Abstract

This paper introduces a task of identifying and semantically classifying lexical expressions in running text. We investigate the online reviews genre, adding semantic supersense annotations to a 55,000 word English corpus that was previously annotated for multiword expressions. The noun and verb supersenses apply to full lexical expressions, whether single- or multiword. We then present a sequence tagging model that jointly infers lexical expressions and their supersenses. Results show that even with our relatively small training corpus in a noisy domain, the joint task can be performed to attain 70% class labeling F_1.

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. A more portable and scalable—yet still linguistically-grounded—way to represent lexical meanings is with coarse-grained semantic classes. Here we build on prior work with an inventory of semantic classes (for nouns and verbs) known as **supersenses**. The 41 supersenses resemble the types used for named entities (PERSON, LOCATION, etc.), but are more general, with semantic categories relevant to common nouns and verbs as well. As a result, their application to sentences is dense (describing a large proportion of tokens), in contrast to annotations that only describe named entities.

Because most supersense tagging studies have worked with data originally annotated for fine-grained WordNet senses, then automatically mapped to supersenses, the resulting systems have been tied to the lexical coverage of WordNet. Schneider et al. (2012) and Johannsen et al. (2014) overcame this limitation in part by annotating supersenses directly in text; thus, nouns and verbs not in WordNet were not neglected. However, the issue of which *units* ought to receive supersenses has not been addressed satisfactorily. We argue that the semantically holistic nature of **multiword expressions** (MWEs) including idioms, light verb constructions, verb-particle constructions, and many compounds (Baldwin and Kim, 2010) means that they should be considered as units for manual and automatic supersense tagging.

Below, we motivate the need for an integrated representation for broad-coverage lexical semantic analysis that identifies MWEs and labels single- and multiword noun and verb expressions with supersenses (§2). By annotating supersenses directly on sentences with existing comprehensive MWE annotations, we circumvent WordNet's spotty coverage of many kinds of MWEs (§3). Then we demonstrate that the two kinds of information are readily combined in a discriminative sequence tagging model (§4). Notably, our analyzer handles gappy expressions that are ignored by existing supersense taggers, and it marks miscellaneous MWEs even though they do not receive a noun or verb supersense.

Our annotations of the REVIEWS section of the English Web Treebank (Bies et al., 2012), which

Human Language Technologies: The 2015 Annual Conference of the North American Chapter of the ACL, pages 1537–1547,
Denver, Colorado, May 31 – June 5, 2015.

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 §3.2)		
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.

enrich the MWE annotations of the CMWE corpus[1] (Schneider et al., 2014b), are publicly released under the name STREUSLE.[2] This includes new guidelines for verb supersense annotation. Our open-source tagger, implemented in Python, is available from that page as well.

2 Background: Supersense Tags

WordNet's **supersense** categories are the top-level hypernyms in the taxonomy (sometimes known as **semantic fields**) which are designed to be broad enough to encompass all nouns and verbs (Miller, 1990; Fellbaum, 1990).[3]

The 26 noun and 15 verb supersense categories are listed with examples in table 1. Some of the names overlap between the noun and verb inventories, but they are to be considered separate categories; hereafter, we will distinguish the noun and verb categories with prefixes, e.g. N:COGNITION vs. V:COGNITION.

Though WordNet synsets are associated with lexical entries, the supersense categories are unlexicalized. The N:PERSON category, for instance, contains synsets for both *principal* and *student*. A different sense of *principal* falls under N:POSSESSION.

As far as we are aware, the supersenses were originally intended only as a method of organizing the WordNet structure. But Ciaramita and Johnson (2003) pioneered the coarse word sense disambiguation task of **supersense tagging**, noting that the supersense categories provided a natural broadening of the traditional named entity categories to encompass all nouns. Ciaramita and Altun (2006) later expanded the task to include all verbs, and applied a supervised sequence modeling framework adapted from NER. Evaluation was against manually sense-tagged data that had been automatically converted to the coarser supersenses. Similar taggers have since been built for Italian (Picca et al., 2008) and Chinese (Qiu et al., 2011), both of which have their own WordNets mapped to English WordNet.

Although many of the annotated expressions in existing supersense datasets contain multiple words, the relationship between MWEs and supersenses has not received much attention. (Piao et al. (2003, 2005) did investigate MWEs in the context of a lexical tagger with a finer-grained taxonomy of semantic classes.) Consider these examples from online reviews:

(1) IT IS NOT A HIGH END STEAK HOUSE

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

HIGH END functions as a unit to mean 'sophisticated, expensive'. (It is not in WordNet, though

[1] `http://www.ark.cs.cmu.edu/LexSem/`

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

[3] WordNet synset entries were originally partitioned into **lexicographer files** for these coarse categories, which became known as "supersenses." The `lexname` function in WordNet/NLTK (Bird et al., 2009) returns a synset's lexicographer file.

A subtle difference is that a special file called `noun.Tops` contains each noun supersense's root synset (e.g., `group.n.01` for N:GROUP) as well as a few miscellaneous synsets, such as `living_thing.n.01`, that are too abstract to fall under any single supersense. Following Ciaramita and Altun (2006), we treat the latter cases under an N:OTHER supersense category and merge the former under their respective supersense.

it could be added in principle.) Assigning a semantic class such as N:LOCATION to *END* would, in our judgment, be overly literal. To paint a coherent picture of the meaning of this sentence, it is better to treat *HIGH END* as a single unit, and because it serves as an adjective rather than a noun or verb, leave it semantically unclassified.[4]

STEAK HOUSE is arguably an entrenched enough compound that it should receive a single supersense—in fact, WordNet spells it without a space. The phrases *white pages*, *high school*, *(get) in touch (with)*, *track... down*, and *one by one* all are listed as MWEs in WordNet. As detailed in §4.1 below, the conventional BIO scheme used in existing supersense taggers is capable of representing most of these. However, it does not allow for gappy (discontinuous) uses of an expression, such as ***track*** *people* ***down***.

The corpus and analyzer presented in this work address these shortcomings by integrating a richer, more comprehensive representation of MWEs in the supersense tagging task.

3 Supersense Annotation for English

As suggested above, supersense tags offer a practical semantic label space for an integrated analysis of lexical semantics in context. For English, we have created the STREUSLE dataset, which fully annotates the REVIEWS corpus (55k words) for noun and verb supersenses in a manner consistent with Schneider et al.'s (2014b) multiword expression annotations.

Schneider et al. (2012) offered a methodology for noun supersense annotation in Arabic Wikipedia, and predicted that it would port well to other languages and domains. Our experience with English web reviews has borne this out. We generally adhered to the same supersense annotation process (for nouns); the most important difference was that the data had already been annotated for MWEs, and supersense labels apply to any strong[5] MWEs as a whole.

The same annotators had already done the MWE annotation; whenever they encountered an apparent mistake from an earlier stage (usually an oversight), they were encouraged to correct it. Our annotation interface supports modification of MWEs as well as supersense labels in one view.

To lessen the cognitive burden when reasoning about tagsets, supersense annotation was broken into separate phases: first we annotated nearly the entire REVIEWS corpus for noun supersenses; then we made another pass to annotate for verbs. Roughly a tenth of the sentences were saved for a combined noun+verb phase at the end; annotators reported that constantly switching their attention between the two tagsets made this mode of annotation more difficult.

3.1 Nouns

Targets. Per the annotation standard, all noun singletons and noun-like MWEs should receive a noun supersense label. Annotation targets were determined heuristically from the gold (PTB-style) POS tags in the corpus: all lexical expressions containing a noun[6] were selected. This heuristic overpredicts noun-like MWEs occasionally because it does not check the syntactic status of the MWE as a whole. During this phase, the backtick symbol (`) was therefore reserved for MWEs (such as light verb constructions) that contain a noun but should not receive a noun supersense.[7] The annotation interface prevented submission of blank annotation targets to avoid oversights.

Tagset conventions. Several brief annotation rounds were devoted to practice with Schneider et al.'s (2012) noun annotation guidelines,[8] since the annotators were new to the scheme. Metonymy posed the chief difficulty in this domain: institutions with a premises (such as restaurants, hotels, and schools) are frequently ambiguous between N:GROUP (institution as a whole), N:ARTIFACT (the building), and N:LOCATION (site as a whole). Our convention was to use the reading that seemed most salient in context: for example, *restaurant* in a comment about the qual-

[4] Future supersense annotation schemes for additional parts of speech could be assimilated into our framework. Tsvetkov et al. (2014) take a step in this direction for adjectives.

[5] The CMWE corpus distinguishes **strong** and **weak** MWEs—essentially, the former are strongly entrenched and likely non-compositional, whereas weak MWEs are merely statistically collocated. See Schneider et al. (2014b) for details. Because they are deemed semantically compositional, weak MWEs do not receive a supersense as a whole.

[6] Specifically, any POS tag starting with N or ADD (web addresses); pronouns were excluded.

[7] Pronouns like *anything* also fall into this category because they are POS-tagged as nouns.

[8] http://www.ark.cs.cmu.edu/ArabicSST/corpus/guidelines.html

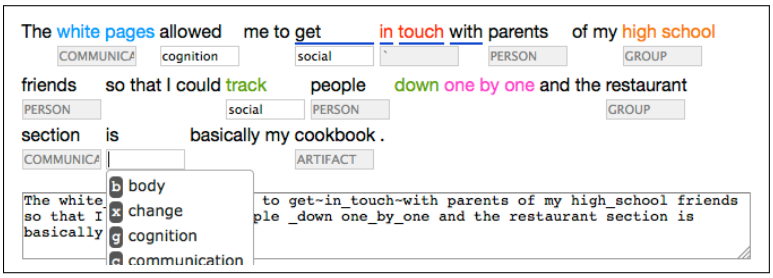

Figure 1: Annotation interface, with dropdown menu for verb supersenses. The large text box at the bottom can be used to edit the MWE annotation by typing underscores and tildes to connect tokens.

ity of the service would be labeled N:GROUP.[9] Some subjectivity is involved, suggesting that the scheme is not ideal for such multifaceted concepts.

3.2 Verbs

Targets. The set of lexical expressions that should receive a verb supersense consists of (a) all verb singletons that are not auxiliaries, and (b) all verb-like MWEs. Again, simple but overly liberal heuristics were used to detect annotation targets, so wherever the heuristics overpredicted, annotators entered:

- `a for auxiliary verbs
- `j for adjectives (some *-ing* and *-ed* adjectives are POS-tagged as VBG and VBD, respectively)
- ` for all other cases

Tagset conventions. We wrote new guidelines to characterize the verb supersenses for annotation. They briefly define and exemplify each category, and also relate them via precedence rules: e.g., the rule

$$\{\text{V:PERCEPTION}, \text{V:CONSUMPTION}\} > \text{V:BODY} > \text{V:CHANGE}$$

stipulates that verbs of perception or consumption (*hear*, *eat*, etc.) be labeled as such rather than the less specific class V:BODY. The precedence rules help to resolve many of the cases of meaning overlap between the categories. The guidelines were developed over several weeks and informed by annotation difficulties and disagreements. We release them along with the STREUSLE corpus.

3.3 Interface

We extended the online MWE annotation tool of Schneider et al. (2014b) to also support supersense labeling, as well as grouping tokens into multiword lexical expressions. This is visualized in figure 1. Specifically, singletons and strong MWEs may receive labels (subject to a POS filter). This allows the two types of annotation to be worked on in tandem, especially when a supersense annotator wishes to change a multiword grouping. The tool offers an autocomplete dropdown menu when typing a tag name, and validates that the submitted annotation is complete and internally consistent. Additionally, the tool provides a complete version history of the sentence and a "reconciliation" mode that merges two users' annotations of a sentence, flagging any differences for manual resolution; these features are extremely useful when breaking the annotation down into multiple rounds among several annotators.

3.4 Quality Control

There were 2 primary annotators and 3 others who participated in annotation to a lesser degree, including the first author of this paper, whose role was mainly supervisory. All 5 hold bachelor's degrees in linguistics. The annotators were trained in the noun supersense annotation scheme of Schneider et al. (2012) and cooperatively developed and documented interpretations for the verb supersenses. Our main quality control mechanism for the annotation process was to obtain two independent annotations for every sentence—differences between them were reconciled by negotiation (between the two annotators in most cases, and between the two primary annotators in a small number of cases).

To get a sense of the difficulty of the task, we examine the annotation history for a sample of sentences to measure inter-annotator agreement. Estimated between the 2 primary annotators on the batch of sentences annotated last during each phase (350, 302, and 379 sentences, respectively), inter-annotator F_1 scores (excluding auxiliaries and other miscellaneous categories) are: 76% for noun expression supersenses after the noun phase, 93% for verb expression supersenses after the verb phase, and 88% for all supersenses after the combined annotation phase.[10] These

[9] This rule is sometimes at odds with WordNet, which only lists N:ARTIFACT for *hotel* and *restaurant*.

[10] Cohen's κ, limited to tokens for which both annotators

are over different sentences, so they are not directly comparable, but they point to the robustness of the annotation scheme. Thanks to the double annotation plus reconciliation procedure, these numbers should underestimate the reliability of the final annotations.

3.5 Corpus Statistics

A total of 9,000 noun mentions (1,300 of them MWEs) and 7,800 verb mentions (1,200 MWEs) incorporating 20,000 word tokens are annotated.[11] Table 1 shows supersense mention counts and the most frequent example of each category in the corpus.

3.6 Copenhagen Supersense Data

An independent English noun+verb supersense annotation effort targeting the Twitter domain was undertaken by the COASTAL lab at the University of Copenhagen (Johannsen et al., 2014). The overarching goal of annotating supersenses directly in running text was the same as in the present work, but there are three important differences. First, general-purpose MWE annotations were not considered in that work; second, sentences were pre-annotated by a heuristic system and then manually corrected, whereas here the annotations are supplied from scratch; and third, Johannsen et al. (2014) provided minimal instructions and training to their annotators, whereas here we have worked hard to encourage consistent interpretations of the supersense categories. Johannsen et al. have released their annotations on two samples of tweets (over 18,000 tokens in total).

Johannsen et al.'s dataset illustrates why supersense annotation by itself is not the same as the full scheme for lexical semantic analysis proposed here. Many of the expressions that they have supersense-annotated as single-word nouns/verbs probably would have been part of larger units in MWE annotation: examining Johannsen et al.'s in-house sample, multiword chunks arguably should have been used for verb phrases like *gain entry*, *make sure*, and *make it* ('succeed'), and for verb-particle constructions like *take over*, *find out*, and *check out* ('ogle'). Moreover, in the traditional supersense annotation scheme, there are no chunks not labeled with a supersense; thus, e.g., PPs such as *on tap*, *of ALL-Time*, and *up to [value limit]* are not chunked.

Many of the nominal expressions in Johannsen et al.'s (2014) data appear to have overly liberal boundaries, grouping perfectly compositional modifiers along with their heads as a multiword chunk: e.g., *Panhandling Ban*, *Panda Cub*, *farm road crash*, and *Tomic's dad*. Presumably, some of these were boundary errors made by the heuristic pre-annotation system that human annotators failed to notice.

4 Automatic Tagging

We now turn to automating the combined multiword expression and supersense prediction task in a single statistical model.

4.1 Background: Supersense Tagging with a Discriminative Sequence Model

Ciaramita and Altun's (2006) model represents the state of the art for full[12] English supersense tagging on the standard SemCor test set, achieving an F_1 score of 77%. It is a feature-based discriminative sequence model learned in a supervised fashion with the structured perceptron (Collins, 2002).

For Ciaramita and Altun (2006) and hereafter, sequences correspond to sentences, with each sentence pre-segmented into words according to some tokenization. Figure 2 shows how token-level tags combine Ramshaw and Marcus (1995)–style BIO flags with supersense class labels to represent the segmentation and supersense labeling of a sentence. These tags are observed during training, predicted at test time, and compared against the gold standard tags.

Ciaramita and Altun's (2006) model uses a simple feature set capturing the lemmas, word shapes, and parts of speech of tokens in a small context window, as well as the supersense category of the first WordNet sense of the current word. (WordNet senses are ordered roughly by frequency.) On SemCor data, the model achieves a 10% absolute improvement in F_1 over the first sense baseline.

assigned a supersense, is very similar: .76, .93, and .90, respectively, reflecting strong agreement.

[11] This excludes 1,200 auxiliary verb mentions, 100 of which are MWEs: *have to*, *is going to*, etc.

[12] Paaß and Reichartz (2009) train a similar sequence model for classifying noun and verb supersenses, but treat multiword phrases as single words. Their model is trained as a CRF rather than a structured perceptron, and adds LDA word cluster features, but the effects of these two changes are not separated in the experiments. They also find benefit from constraining the label space according to WordNet for in-vocabulary words (with what they call "lumped labels").

United States financier and philanthropist (1855 - 1937)
$_B$N:LOCATION $_I$N:LOCATION $_B$N:PERSON O $_B$N:PERSON O $_B$N:TIME O $_B$N:TIME O

Figure 2: A supersense tagging shown with per-token BIO tags in the style of Ciaramita and Altun (2006).

The white pages allowed me to get in touch with parents of my high school
$_B$N:COMMUNICATION Ī $_0$V:COGNITION 0 0 $_B$V:SOCIAL Ĩ Ī Ĩ $_0$N:PERSON 0 0 $_B$N:GROUP Ī
friends so that I could track people down one by one
$_0$N:PERSON 0 0 0 0 $_B$V:SOCIAL $_0$N:PERSON Ī B Ī Ī

Figure 3: Tagging for part of the lexical semantic analysis depicted in figure 1. Note that for nominal and verbal MWEs, the supersense label is only attached to the first tag of the expression.

Though our focus in this paper is on English, automatic supersense tagging has also been explored in Italian (Picca et al., 2008, 2009; Attardi et al., 2010, 2013; Rossi et al., 2013), Chinese (Qiu et al., 2011), and Arabic (Schneider et al., 2013).

4.2 Model

Like Ciaramita and Altun (2006) and Schneider et al. (2014a), we train a first-order structured perceptron (Collins, 2002) with averaging. This is a standard discriminative modeling setup, involving: a linear scoring function over features of input–output pairs; a Viterbi search to choose the highest-scoring valid output tag sequence given the input; and an online learning algorithm that makes M passes through the training data, searching for the best tagging given the current model and updating the parameters (linear feature weights) where the best tagging doesn't match the gold tagging. With a first-order Markov assumption and tagset $\mathcal{Y}$, the Viterbi search for a sentence $\mathbf{x}$ requires $O(|\mathcal{Y}|^2 \cdot |\mathbf{x}|)$ runtime. The dataset used to train and evaluate the model, the tagging scheme, and the features are described below.

4.3 Data

The STREUSLE dataset, as described in §3, is annotated for multiword expressions as well as noun and verb supersenses and auxiliary verbs. We use this dataset for training and testing an integrated lexical semantic analyzer. Schneider et al. (2014a) used the CMWE dataset—i.e., the same REVIEWS sentences, but annotated only for MWEs. A handful of apparent errors in the MWE analyses were fixed in the course of our supersense annotation.

4.4 Tagset

In the STREUSLE dataset, supersense labels apply to *strong* noun and verb expressions—i.e., singleton nouns/verbs as well as strong nominal/verbal MWEs. Weak MWEs are not holistically labeled with a supersense (see fn. 5).

The 8-way scheme. To encode the lexical segmentation via token-level tags, we use the 8-way scheme from Schneider et al. (2014a) for positional flags. The 8-way scheme extends Ramshaw and Marcus's (1995) BIO chunking tags to also encode (a) a strong/weak distinction for MWEs, and (b) gappy MWEs (there is no formal limit on the number of gaps per MWE or the number of other lexical expressions occurring within each gap, though there is a limit of one level of nesting). The 4 lowercase positional flags indicate that an expression is within a gap, and otherwise have the same interpretation as their uppercase counterparts, which are:

- `0` for single-word expressions
- `B` for the first word of an MWE
- `Ī` for a word continuing a *strong* MWE
- `Ĩ` for a word weakly linked to its predecessor, forming a *weak* MWE[13]

As with the original BIO scheme, a globally well-formed sequence of tags in the 8-tag scheme can be constructed by respecting bigram constraints.[14]

Adding class labels. The tagset used to annotate the data for our tagger combines 8-way positional flags with supersense class labels. We decorate class labels only on beginners of strong lexical expressions—so this includes `0` or `o` on a single-word noun or verb, but always excludes `Ī` and `ī`.[15] Figure 3

[13] Weak MWE links may join together strong MWEs.

[14] Among these constraints are: `B` must always be immediately followed by `Ī` or `Ĩ` (because `B` marks the beginning of an MWE); and within-gap (lowercase-tagged) tokens must immediately follow a tag other than `0` and precede a tag other than `0` or `B`.

[15] Unlike prior work with the plain BIO scheme, we do not include the class in tags continuing a (strong) MWE, though the

gives an example. In this formulation, bigram constraints are sufficient to ensure a globally consistent tagging of the sentence.

There are $|\mathcal{N}| = 26$ noun supersense classes and $|\mathcal{V}| = 16$ verb classes (including the auxiliary verb class, abbreviated `a). In principle, then, there are

$$\underbrace{|\{\texttt{0 o B b \tilde{I} \tilde{i}}\}|}_{6} \times \underbrace{(1 + |\mathcal{N}| + |\mathcal{V}|)}_{43} + \underbrace{|\{\bar{\texttt{I}}\ \bar{\texttt{i}}\}|}_{2} = 260$$

possible tags encoding position 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 **train**, yielding $|\mathcal{Y}| = 146$.

We also run a condition where the supersense refinements are collapsed, i.e. $\mathcal{Y}$ consists of the 8 MWE tags. This allows us to measure the impact of the supersenses on MWE identification performance.

4.5 Features

We constrast three feature sets for full supersense tagging: (a) Schneider et al.'s (2014a) basic MWE features, which include lemmas, POS tags, word shapes, and whether the token potentially matches entries in any of several multiword lexicons; (b) the basic MWE features plus the Brown clusters (Brown et al., 1992) used by Schneider et al. (2014a); and (c) the basic MWE features and Brown clusters, plus several new features shown in figure 4. Chiefly, these new features consult the supersenses of WordNet synsets associated with words in the sentence: the first WordNet supersense feature is inspired by Ciaramita and Altun (2006) and subsequent work on supersense tagging, while the has-supersense feature is novel. There is also a feature aimed at distinguishing auxiliary verbs from main verbs, and new capitalization features take into account the capitalization of the first word in the sentence and the majority of words in the sentence. To keep the system as modular as possible, we refrain from including any features that depend on a syntactic parser.

class label should be interpreted as extending across the entire expression. This is for a technical reason: as our scheme allows for gaps, the classes of the tags flanking a gap in a strong MWE would be required to match for the analysis to be consistent. To enforce this in a bigram tagger, the within-gap tags would have to encode the gappy expression's class as well as their own, leading to an undesirable blowup in the size of the state space.

New Capitalization Features

1. capitalized $\wedge$ $[\![i = 0]\!]$ $\wedge$ [majority of tokens in the sentence are capitalized]
2. capitalized $\wedge$ $i > 0$ $\wedge$ w_0 is lowercase

Auxiliary Verb vs. Main Verb Feature

3. pos_i is a verb $\wedge$ $[\![pos_{i+1}$ is a verb $\vee$ (pos_{i+1} is an adverb $\wedge$ pos_{i+2} is a verb)$]\!]$

WordNet Supersense Features (unlexicalized)

Let $cpos_i$ denote the coarse part-of-speech of token i: common noun, proper noun, pronoun, verb, adjective, adverb, etc. This feature aims primarily to inform the supersense label on the first token of nominal compounds and light verb constructions, where the "semantic head" is usually a common noun subsequent to the beginning of the expression:

4. subsequent noun's 1st supersense: where $cpos_i$ is a common noun, verb, or adjective, $cpos_i$ $\wedge$ for the smallest $k > i$ such that pos_k is a common noun, the supersense of the first WordNet synset for lemma λ_k—provided there is no intervening verb (j such that $cpos_j$ is a verb and $i < j < k$)

The following two feature templates depend on the tag y_i. Let $flag(y_i)$ denote the positional flag part of the tag (`0`, `B`, etc.) and $sst(y_i)$ denote the supersense class label:

5. 1st supersense:
   - if $flag(y_i) \in \{\texttt{0}, \texttt{o}\}$: the supersense of the first WordNet synset for lemma λ_i
   - else if $cpos_i$ is a verb and there is a subsequent verb particle at position $k > i$ with no intervening verb: the supersense of the first synset for the compound lemma $\langle \lambda_i, \lambda_k \rangle$ (provided that the particle verb is found in WordNet)
   - otherwise: the supersense of the first WordNet synset for the longest contiguous lemma starting at position i that is present in WordNet: $\langle \lambda_i, \lambda_{i+1}, \ldots, \lambda_j \rangle$ $(j \geq i)$
6. has supersense: same cases as the above, but instead of encoding the highest-ranking synset's supersense, encodes whether $sst(y_i)$ is represented in any of the matched synsets for the given lemma. Note that for a given token, this feature can take on different values for different tags.

Figure 4: New features for MWE and supersense tagging. They augment the basic MWE feature set of Schneider et al. (2014a), and are conjoined with the current tag, y_i.

The model's percepts (binary or real-valued functions of the input[16]) can be conjoined with any tag $y \in \mathcal{Y}$ to form a feature that receives its own weight

[16] We use the term **percept** rather than "feature" here to emphasize that we are talking about functions of the input only, rather than input–output combinations that each receive a parameter during learning.

(parameter). To avoid having to learn a model with tens of millions of parameters, we impose a percept cutoff during learning: only zero-order percepts that are active at least 5 times in the training data (with any tag) are retained in the model (with features for all tags). There is no minimum threshold for first-order percepts.[17] The resulting models are of a manageable size: about 4 million parameters with the full tagset.

4.6 Experimental Setup

Our setup mostly echoes that of Schneider et al. (2014a). We adopt their **train** (3312 sentences/ 48k words) vs. **test** (500 sentences/7k words) split, and tune hyperparameters by 8-fold cross-validation on **train**. By this procedure we chose a percept cutoff of 5 to use throughout, and tuned the number of training iterations for each experimental condition (early stopping within each cross-validation fold so as to greedily maximize tagging accuracy on the held-out portion, and averaging the best number of iterations across folds). For simplicity, we use oracle POS tags in our experiments and do not use Schneider et al.'s (2014a) recall-oriented cost function. Experiments were managed with Jonathan Clark's `ducttape` tool.[18]

4.7 Results

Table 2 shows full supersense tagging results, separating the MWE identification performance (measured by link-based precision, recall, and F_1; see Schneider et al., 2014a) from the precision, recall, and F_1 of class labels on the first token of each expression (segments with no class label are ignored).[19] Exact tagging accuracy (last column) is higher because it rewards true negatives, i.e. single-word segments with no nominal or verbal class label (the `0` and `o` tags).

Tag space. The sequence tagging framework makes it simple to model MWE identification jointly with supersense tagging: this is accomplished by packing information about both kinds of output into the tags. But there is always a risk that a larger tag space will impair the model's ability to generalize. By comparing the first two rows of the results, we can see that jointly modeling supersenses along with multiword expressions results in only a minor decrease (<2 F_1 points) in MWE identification performance under the most basic feature set. Further, we see that most of that decrease is recovered with richer features. Thus, we conclude that it is empirically reasonable to model these phenomena together.

Runtime. Our final system (146 tags; last row of table 2) tags ≈140 words (10 sentences) per second.

Features. Comparing the bottom three rows in the table indicates that features that generalize beyond lexical items lead to better supersense labeling. The best model has access to supersense information in the WordNet lexicon; it is 4 F_1 points better at choosing the correct class label than its nearest competitor, which relies on word clusters to abstract away from individual lexical items. Nouns, verbs, and auxiliaries all see improvements.

We also inspect the learned parameters. The highest-weighted parameters suggest that the best model relies heavily on the supersense lookup features, whereas the second-best model—lacking those—in large part relies on Brown clusters (cf. Grave et al., 2013). The auxiliary verb vs. main verb feature in the best model is highly weighted as well, helping to distinguish between `a and V:STATIVE.

Polysemy. We have motivated the task of supersense tagging in part as a coarse form of word sense disambiguation. Therefore, it is worth investigating how well the learned model manages to choose the correct supersense for nouns and verbs that are ambiguous in the data. A handful of lemmas in **test** have at least two different supersenses predicted several times; an examination of four such lemmas in table 3 shows that for three of them the tagging accuracy exceeds the majority baseline. In the case of *look*, the model is usually able to distinguish between V:COGNITION (as in *looking for a company with decent rates*) and V:PERCEPTION (as in *sometimes the broccoli looks browned around the edges*).

Out-of-domain baseline. To assess the importance of in-domain data for learning, we used a SemCor-trained supersense tagger—a reimplemen-

[17] Zero-order percepts are percepts which are to be conjoined with only the present tag to form zero-order features. First-order percepts are to be conjoined with the present and previous tags.

[18] `https://github.com/jhclark/ducttape/`

[19] We count the class label only once for MWEs—otherwise this measure would be strongly dependent on segmentation performance. However, the MWE predictions do have an effect when the prediction and gold standard disagree on which token begins a strong nominal or verbal expression.

				MWE ID			Class labeling						Tag
Feature Set	$\lvert\mathcal{Y}\rvert$	Model Size	M	P	R	F_1	P	R	F_1	NSST R	VSST R	Aux R	Acc
MWE	8	194k	4	**72.97**	**55.55**	**63.01**	—	—	—	—	—	—	—
MWE	146	3,555k	5	67.77	55.76	61.14	64.68	66.78	65.71	59.14	71.64	93.71	80.73
MWE+clusters	146	4,371k	5	68.55	**56.73**	62.04	65.69	67.76	66.71	61.49	71.34	92.45	81.20
MWE+clusters+SST	146	4,388k	4	**71.05**	56.24	**62.74**	**69.47**	**71.90**	**70.67**	**66.95**	**74.17**	**94.97**	**82.49**

Table 2: Results on **test** for lexical semantic analysis of noun and verb supersenses and MWEs with increasingly complex models. Class labeling performance is given in aggregate, and class labeling *recall* is further broken down into noun supersense tagging (NSST), verb supersense tagging (VSST), and auxiliary verb tagging. All of these results use a percept cutoff of 5. The first result row uses a collapsed tagset (just the MWE status) rather than predicting full supersense labels, as described in §4.4. The number of training iterations M was tuned by cross-validation on **train**. The best result in each column and section is bolded.

lemma	unique SSTs	majority baseline	accuracy
get	7 gold, 8 pred.	**12/28**	6/28
look	2 gold, 3 pred.	8/13	**12/13**
take	5 gold, 5 pred.	8/21	**11/21**
time(s)	3 gold, 2 pred.	8/14	**9/14**

Table 3: Four polysemous lemmas and counts of their gold vs. predicted supersenses in **test** (limited to cases where both the gold standard tag and the predicted tag included a supersense). The distribution of gold supersenses for *take*, for example, is V:SOCIAL: 8, V:MOTION: 7, V:POSSESSION: 1, V:STATIVE: 4, V:EMOTION: 1.

tation of Ciaramita and Altun (2006)[20]—to tag our test data in the reviews domain. By our class labeling evaluation, the result is 51.05% precision, 48.93% recall, and 49.97% F_1.[21] Even without word clusters or the supersense-tailored features of figure 4, our simplest in-domain model reaches 65.71% F_1. Though there are minor differences in features between the two models, both are first-order structured perceptron taggers. We believe that this wide gulf is primarily an artifact of the training data. The annotation methodology was very different (direct MWE and supersense annotation in our case, vs. relying on mappings from WordNet synsets in the case of SemCor), and the vocabulary and style are vastly different between casual online writing and edited prose. Building lexical semantic models that are robust to many domains at once will require further experimentation, and in our estimation, additional annotated resources that cover a fuller spectrum of written language.

5 Conclusion

We have integrated the multiword expression identification task formulated in Schneider et al. (2014a) with the supersense tagging task of Ciaramita and Altun (2006). Supersenses offer coarse-grained and broadly applicable semantic labels for lexical expressions and naturally complement multiword expressions in lexical semantic analysis. We have annotated English online reviews for supersenses, including developing detailed annotation criteria for verbs. Experiments with discriminative joint tagging of MWEs and supersenses establish a strong baseline for future work, which may incorporate new features, richer models, and indirect forms of supervision (cf. Grave et al., 2013; Johannsen et al., 2014) for this task. We also expect future investigations will apply our tagger to a downstream task such as semantic parsing or machine translation (for further discussion of potential applications, see Schneider, 2014, pp. 179–189). Our data and open-source software is available at `http://www.ark.cs.cmu.edu/LexSem/`.

Acknowledgments

We thank our energetic annotators, Nora Kazour, Spencer Onuffer, Emily Danchik, and Michael T. Mordowanec, as well as Chris Dyer, Lori Levin, Ed Hovy, Tim Baldwin, Mark Steedman, and anonymous reviewers for useful feedback on the technical content. This research was supported in part by NSF CAREER grant IIS-1054319 and DARPA grant FA8750-12-2-0342 funded under the DEFT program.

[20] By Michael Heilman (Heilman, 2011, pp. 47–48); downloaded from: `http://www.ark.cs.cmu.edu/mheilman/questions/SupersenseTagger-10-01-12.tar.gz`

[21] Excluding auxiliaries (which are not part of the original supersense representation and thus not predicted by Heilman's tagger) from the evaluation, recall rises to 52.50% and F_1 to 51.76%.

References

Giuseppe Attardi, Luca Baronti, Stefano Dei Rossi, and Maria Simi. 2013. SuperSense Tagging with a Maximum Entropy Markov Model. In Bernardo Magnini, Francesco Cutugno, Mauro Falcone, and Emanuele Pianta, editors, *Evaluation of Natural Language and Speech Tools for Italian*, number 7689 in Lecture Notes in Computer Science, pages 186–194. Springer, Berlin.

Giuseppe Attardi, Stefano Dei Rossi, Giulia Di Pietro, Alessandro Lenci, Simonetta Montemagni, and Maria Simi. 2010. A resource and tool for super-sense tagging of Italian texts. In Nicoletta Calzolari, Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, Mike Rosner, and Daniel Tapias, editors, *Proc. of LREC*, pages 2242–2248. Valletta, Malta.

Timothy Baldwin and Su Nam Kim. 2010. 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, FL.

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

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit*. O'Reilly Media, Inc., Sebastopol, CA.

Peter F. Brown, Peter V. deSouza, Robert L. Mercer, Vincent J. Della Pietra, and Jenifer C. Lai. 1992. Class-based n-gram models of natural language. *Computational Linguistics*, 18(4):467–479.

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

Massimiliano Ciaramita and Mark Johnson. 2003. Supersense tagging of unknown nouns in WordNet. In Michael Collins and Mark Steedman, editors, *Proc. of EMNLP*, pages 168–175. Sapporo, Japan.

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

Christiane Fellbaum. 1990. English verbs as a semantic net. *International Journal of Lexicography*, 3(4):278–301.

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

Edouard Grave, Guillaume Obozinski, and Francis Bach. 2013. Hidden Markov tree models for semantic class induction. In *Proc. of CoNLL*, pages 94–103. Sofia, Bulgaria.

Michael Heilman. 2011. *Automatic factual question generation from text*. Ph.D. dissertation, Carnegie Mellon University, Pittsburgh, Pennsylvania. URL `http://www.ark.cs.cmu.edu/mheilman/questions/papers/heilman-question-generation-dissertation.pdf`.

Anders Johannsen, Dirk Hovy, Héctor Martínez Alonso, Barbara Plank, and Anders Søgaard. 2014. More or less supervised supersense tagging of Twitter. In *Proc. of *SEM*, pages 1–11. Dublin, Ireland.

George A. Miller. 1990. Nouns in WordNet: a lexical inheritance system. *International Journal of Lexicography*, 3(4):245–264.

Gerhard Paaß and Frank Reichartz. 2009. Exploiting semantic constraints for estimating supersenses with CRFs. In *Proc. of the Ninth SIAM International Conference on Data Mining*, pages 485–496. Sparks, Nevada.

Scott S. L. Piao, Paul Rayson, Dawn Archer, Andrew Wilson, and Tony McEnery. 2003. Extracting multiword expressions with a semantic tagger. In *Proc. of the ACL 2003 Workshop on Multiword Expressions: Analysis, Acquisition and Treatment*, pages 49–56. Sapporo, Japan.

Scott Songlin Piao, Paul Rayson, Dawn Archer, and Tony McEnery. 2005. Comparing and combining a semantic tagger and a statistical tool for MWE extraction. *Computer Speech & Language*, 19(4):378–397.

Davide Picca, Alfio Massimiliano Gliozzo, and Simone Campora. 2009. Bridging languages by SuperSense entity tagging. In *Proc. of NEWS*, pages 136–142. Suntec, Singapore.

Davide Picca, Alfio Massimiliano Gliozzo, and Massimiliano Ciaramita. 2008. Supersense Tagger for Italian. In Nicoletta Calzolari, Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odjik, Stelios Piperidis, and Daniel Tapias, editors, *Proc. of LREC*, pages 2386–2390. Marrakech, Morocco.

Likun Qiu, Yunfang Wu, Yanqiu Shao, and Alexander Gelbukh. 2011. Combining contextual and structural information for supersense tagging of Chinese unknown words. In *Computational Linguistics and Intelligent Text Processing: Proceedings of the 12th International Conference (CICLing'11)*, volume 6608 of *Lecture Notes in Computer Science*, pages 15–28. Springer, Berlin.

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

Stefano Dei Rossi, Giulia Di Pietro, and Maria Simi. 2013. Description and results of the SuperSense tagging task. In Bernardo Magnini, Francesco Cutugno, Mauro Falcone, and Emanuele Pianta, editors, *Evaluation of Natural Language and Speech Tools for Italian*, number 7689 in Lecture Notes in Computer Science, pages 166–175. Springer, Berlin.

Nathan Schneider. 2014. *Lexical Semantic Analysis in Natural Language Text*. Ph.D. dissertation, Carnegie Mellon University, Pittsburgh, Pennsylvania, USA. URL `http://www.cs.cmu.edu/~nschneid/thesis/thesis-print.pdf`.

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

Nathan Schneider, Behrang Mohit, Chris Dyer, Kemal Oflazer, and Noah A. Smith. 2013. Supersense tagging for Arabic: the MT-in-the-middle attack. In *Proc. of NAACL-HLT*, pages 661–667. Atlanta, Georgia, USA.

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

Nathan Schneider, Spencer Onuffer, Nora Kazour, Emily Danchik, Michael T. Mordowanec, Henrietta Conrad, and Noah A. Smith. 2014b. 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.

Yulia Tsvetkov, Nathan Schneider, Dirk Hovy, Archna Bhatia, Manaal Faruqui, and Chris Dyer. 2014. Augmenting English adjective senses with supersenses. 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 4359–4365. Reykjavík, Iceland.