Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence

English Light Verb Construction Identification Using Lexical Knowledge

Wei-Te Chen
Department of Computer Science
University of Colorado at Boulder
Weite.Chen@colorado.edu

Claire Bonial
Department of Linguistics
University of Colorado at Boulder
Claire.Bonial@colorado.edu

Martha Palmer
Department of Linguistics
University of Colorado at Boulder
Martha.Palmer@colorado.edu

Abstract

This research describes the development of a supervised classifier of English light verb constructions, for example, *take a walk* and *make a speech*. This classifier relies on features from dependency parses, OntoNotes sense tags, WordNet hypernyms and WordNet lexical file information. Evaluation shows that this system achieves an 89% F_1 score (four points above the state of the art) on the BNC test set used by Tu & Roth (2011), and an F_1 score of 80.68 on the OntoNotes test set, which is significantly more challenging. We attribute the superior F_1 score to the use of our rich linguistic features, including the use of WordNet synset and hypernym relations for the detection of previously unattested light verb constructions. We describe the classifier and its features, as well as the characteristics of the OntoNotes light verb construction test set, which relies on linguistically motivated PropBank annotation.

1 Introduction

As one construction in which the relational semantics of verbs can be extended in novel ways, English light verb constructions (LVCs) represent a powerfully expressive resource of English. These constructions, such as *make an offer* and *give a groan*, are thought to consist of a semantically general verb and a noun that denotes an event or state. However, an exact definition of an LVC, which would precisely delimit these constructions from either idiomatic expressions or compositional, 'heavy' usages of the same verbs (e.g. *She made a dress; He gave me a present;* etc.) remains under debate. As a result, it is no surprise that the automatic detection of English LVCs also remains challenging, especially given the semi-productive nature of LVCs, which allows for novel LVCs to enter the language. Novel LVCs are particularly difficult to detect, yet it is key for Natural Language Processing (NLP) systems to identify and interpret LVCs correctly by recognizing, for example, that *She made an offer to buy the house for $1.5 million* is an *offering* event, rather than a *making* or *creation* event.

This research describes a system for the automatic detection of LVCs. Related work is given in Section 2, and relevant linguistic resources are described in Section 3. A comparison of the OntoNotes LVC annotations to the British National Corpus[1] LVC annotations, used by Tu and Roth (2011), is given in Section 4. The description of our system is presented in Section 5, and our experiments and results in 6, followed by concluding remarks and future work.

2 Related Work

There are two main approaches for the automatic identification of LVCs: contextually-based and statistically-based. Contextually-based approaches detect the surrounding tokens and decide whether the verb, noun pair with these context words should be considered an LVC. Vincze et al. (2003) propose a contextually-based model, with a conditional random fields machine learning method, for detecting English and Hungarian LVCs. Evaluation showed that their model performs well in various domains of LVCs and performs well in detecting low-frequency LVCs. On the other hand, the statistically-based approach finds LVCs among verb, noun pairs from a well-defined set of verbs and eventive nouns (nouns denoting events, like *declaration*), then a classifier function decides if a pair is an LVC or not. Van de Cruys and Moirn (2007) propose a statistically and semantically-based method for recognizing verb-preposition-noun dependency relation combinations of LVCs. Furthermore, Gurrutxaga and Alegria (2012) detect idiomatic and light verb-noun pairs from Basque, using statistical methods.

To compare these two approaches, Tu and Roth (2011) proposed a Support Vector Machine (SVM) based classifier to identify LVCs. They developed their system using both contextual and statistical features and analyzed the deep interaction between them. They concluded that local contextual features perform better than statistical features on ambiguous examples, and combining them did not give better performance. We also focus on contextual features and find additional features that improve performance.

3 Resources

This research uses several resources: PropBank (PB) (Palmer, Guildea, and Kingsbury 2005), the OntoNotes (ON) sense groupings (Pradhan et al. 2007), WordNet (WN)

[1]http://www.natcorp.ox.ac.uk/XMLedition/

(Fellbaum, Grabowski, and Landes 1998) and the British National Corpus (BNC).

3.1 PropBank

The primary goal of PB was the development of an annotated corpus to be used as training data for supervised machine learning systems. The first PB release consists of 1M words of the Wall Street Journal portion of the Penn Treebank II (Marcus, Santorini, and Marcinkiewicz 1994), annotated with predicate-argument structures for verbs, using semantic role labels for each verb argument. Although the semantic role labels are purposely chosen to be quite generic and theory neutral, Arg0, Arg1, etc., they are still intended to consistently annotate the same semantic role across syntactic variations (Arg0 and Arg1 do consistently correspond to Dowty's (1991) concepts of Proto-Agent and Proto-Patient respectively). For example, the Arg1 or Patient in *John broke the window* is the same window that is annotated as the Arg1 in *The window broke*, even though it is the syntactic subject in one sentence and the syntactic object in the other. Thus, the main goal of PB is to supply consistent, simple, general purpose labeling of semantic roles for a large quantity of coherent text to support the training of automatic semantic role labelers, as the Penn Treebank has supported the training of statistical syntactic parsers.

PB provides a lexical entry for each broad meaning of every annotated verb, including the possible arguments of the predicate and their labels (its 'roleset') and all possible syntactic realizations.[2] This lexical resource is used as a set of verb-specific guidelines for annotation. In addition to numbered roles, PB defines several more general (ArgM, 'Argument Modifier') roles that can apply to any verb, such as LOCation, TeMPoral, and DIRection, etc.

In the past, PB annotation had been restricted to verb relations, but recent work has extended coverage to noun relations and complex relations like LVCs. In current practices, annotators identify light verbs and the main noun predicate in an initial verb pass of annotation. In a second pass, annotation is completed for the full span of the complex predicate, using the roleset of the noun. Consider the example, *Yesterday*-ARGM-TEMPORAL, *John*-ARG0 *made*-REL *an offer*-REL *[to buy the house]*-ARG1 *[for $350,000]*-ARG2, which uses the *offer* roleset:

Arg0: entity offering
Arg1: commodity, thing offered
Arg2: price
Arg3: benefactive or entity offered to

PB ensures that the complete argument structure of the complex predicate receives annotation, regardless of whether the argument is within the domain of locality of the noun or verb, and ensures that the roles assigned reflect the event semantics of the noun.

[2]The PB lexicon of rolesets can be found here: http://verbs.colorado.edu/propbank/framesets-english/

Name	Nouns denoting...
noun.act	acts or actions
noun.cognition	cognitive process
noun.communication	communicative process
noun.event	natural events
noun.feeling	feelings and emotions
noun.location	spatial position
noun.motive	goals
noun.phenomenon	natural phenomena
noun.possession	possession and transfer
noun.process	natural processes
noun.relation	relations between things
noun.state	stable states of affairs

Table 1: WordNet lexical file information types of interest for eventive and stative nouns

3.2 WordNet

WN is a large electronic database of English words,[3] which was in part inspired by work in psycholinguistics investigating how and what type of information is stored in the human mental lexicon (Miller 1995). WN is divided firstly into syntactic categories: nouns, verbs, adjectives and adverbs, and secondly by semantic relations. The semantic relations that organize WN are: synonymy (given in the form of 'synsets'), antonymy, hyponymy (e.g. a Maple is a tree; therefore, *tree* is a hypernym of *Maple*), and meronymy (part-whole relations). These relations make up a complex network of associations that is both useful for computational linguistics and NLP, and also informative in situating a word's meaning with respect to others.

Of particular interest for this research are the synsets, the hyponymic relations of nouns in WN, and the noun's 'type,' as indicated by the lexical file information. For each noun in WN, lexicographers have coded the noun with one primary superordinate, or lexical file, given forty-five numbered options. In our research, nouns that can possibly denote events or states are the focus, because it is these nouns that can theoretically combine with a light verb to form an LVC. The type designations that may denote eventive and stative nouns are listed in Table 1. The use of synset, hyponym and lexical file information (or noun 'type,') is described in Section 5.2.

3.3 OntoNotes

The ON corpus integrates several layers of different annotation types in a single corpus, making it ideal training data for semantic analysis (Pradhan et al. 2007). The five layers of annotation include: 1) the syntactic parse from the Penn Treebank, 2) proposition structure from PB, 3) coarse-grained word senses from the ON sense grouping inventory, 4) named entity types, and 5) anaphoric coreference. The latest release, ON 4.99 (Weischedel et al. 2011), contains 2.6 million English words. In this research, the PB and word sense layers are of primary interest, the latter is described next.

[3]http://wordnet.princeton.edu/wordnet/

The ON sense groupings can be thought of as a more coarse-grained view of WN senses because these sense groupings were based on WN senses, which were successively merged into more coarse-grained senses, based on the results of inter-annotator agreement (Duffield et al. 2007). Essentially, where two annotators were consistently able to distinguish between two senses, the distinction was kept. Where annotators were not able to consistently distinguish between two senses, the senses were reorganized and tagged again. It was found that sense distinctions with this level of granularity can be detected automatically at 87-89% accuracy, making them effective for NLP applications (Dligach and Palmer 2011). This sense inventory was used to annotate ON verbs and nouns with more than three WN senses. Unfortunately, the sense tagging is not complete for all of the ON corpus: there are about one million verbs and nouns in ON 4.99, but only 288,217 of these have sense tags (although many are surely monosemous), including 120,400 nouns with sense tags. Each ON sense also lists which WN senses it includes, providing a mapping between ON annotations and WN senses.

4 Comparison of Light Verb Construction Resources

The existing state of the art system for LVC recognition is arguably that of Tu and Roth (2011), who achieve 86.3% accuracy. To best compare our work to the state of the art, a detailed comparison was made of the resources used by Tu and Roth and the PB LVC data set used in the present work. Tu and Roth construct a dataset of 2,162 English sentences with LVCs drawn from the British National Corpus (BNC). Their approach in constructing this data set differs from that of PB in several ways, and therefore results in resources containing some overlapping and some distinct constructions. Firstly, the authors restrict their annotations to LVCs involving the six most frequent light verbs: *do, get, give, have, make, take*. In the PB annotation process, it is possible for annotators to mark any verb as a light verb, resulting in a corpus that contains seven LVC types with verbs not included in the Tu and Roth data, such as textitbring charges against... and *conduct repairs*. Secondly, Tu and Roth filter their data set by including only LVCs with nouns that are zero-derived nominals (e.g. *offer*), or derivationally related to a verb (e.g. *destruction*). The PB corpus includes an additional 25 LVC types (not found in the Tu and Roth data), which involve nouns that have no etymologically related verb counterpart, such as *take a trip*. Although the PB procedure allows for more variety, this has not resulted in a broader data set with more unique LVC types overall. The comparison shows that there are 115 LVC types that appear in both data sets, 245 LVC types that appear only in the BNC, and 218 LVC types that appear only in ON.

Although the majority of these different types simply arise from the differing sources and genres, there are notably more instances of LVCs involving *get* and *give* in the BNC data set. PB has previously treated many of these usages as 'semi-light,' and opted to annotate both the argument structure of the verb and that of the noun in distinct annotation passes

		have	**do**	**make**	**take**
Token Freq	+	221	128	996	317
	-	1,781	405	671	766
Type Freq	+	62	55	187	46
	-	521	155	249	274
Overlap		47	21	98	27
Portion		75.8%	38.2%	52.4%	58.7%

Table 2: Token and type frequency of positive LVC and non-LVC examples, and the number of overlapping *verb+noun* types that can be either positive or negative examples in ON.

instead of marking these as LVCs. As a result, the BNC data set includes 83 additional types with *get* and *give*. The PB practice has led to some inconsistencies; for example, *give a speech (to the audience)* has been treated as semi-light (since the audience can be seen as either the Recipient of the *give* event or the *speaking* event), while *make a speech* has been treated as an LVC (since there is no similar overlap in the noun and verb relations' roles). To remedy such inconsistencies, PB will be loosening the annotation requirements and including such semi-light usages in the LVC annotations. Table 2 gives an overview of the number of positive LVC types and tokens, and the number of non-LVC tokens and non-LVC *verb+noun* types involving several common light verbs in the ON corpus. The 'overlap' portion indicates the number and percentage of nouns that appear in both positive and negative examples. Notably, this overlap is highest for *have*, indicating that it involves a high number of surface-identical LVC and non-LVC usages.

In summary, the two LVC lexical resources differ in ways that likely provide an advantage for the Tu and Roth system: the BNC data has less variety in the types of nouns that can be involved in LVCs, and it embraces a more general definition of LVCs since it has not distinguished light and semi-light usages.

5 Light Verb Construction Recognition

Our LVC identifier determines what combinations of potential light verbs and eventive nouns should be labeled as LVCs. For a given dependency tree T, the system first checks if T meets certain criteria in order to decide if T should be put into the candidate set (these criteria are described in Section 5.1). Next, the light (or 'Support') verb V_S and eventive noun N_E pair is submitted to an LVC binary classifier, which labels the V_S-N_E pair as an LVC or not an LVC. This supervised classifier is trained with the LibLinear (Fan et al. 2008) algorithm.

5.1 Candidate Identification

The first step for LVC recognition is to select the candidate dependency trees for the training of the classifier. Here, the PB layer of the ON 4.99 corpus is used as a starting point. For this research, we chose to exploit LVCs that are composed of a limited set of the six most frequent light verbs in the data, since these cover 99.26% of the V_S-N_E pairs. In the future, we plan to expand our verb set.

Relation Type	Numbers	Portion
I. Direct Chlid	1,546	87.44%
II. Quantity	16	0.90%
III. Head of the clause	178	10.07%
Sub-Total	**1,740**	**98.42%**
Other Type	**28**	**1.58%**
Total	**1,768**	**100.00%**

Table 3: Distribution of Dependency Relation Type of LVCs in ON 4.99 data

The second step is to select the eventive nouns on which to focus. Our starting point for this process was to make use of a list of eventive and stative nouns drawn from WN (initial list provided by Christiane Fellbaum, personal communication). To begin, only the dependency trees containing the previously mentioned six light verbs and eventive nouns on the candidate list are selected.

The last step is to extract the dependency relation between the light verb V_S and eventive noun N_E. The following three cases are considered:

(i) The N_E is the direct child of the V_S, with the object relation of the V_S (e.g. *They took a look at the trials and tribulations of Martha Stewart.*)

(ii) The N_E follows a quantity or partitive expression, and the quantity is the direct object of the V_S (e.g. *We don't take much of a break for the holidays.*)

(iii) The N_E is the head word of the clause (e.g. *This is very much a statement that the president was not going to be deterred from making.*)

Only the dependency trees containing the V_S and N_E in one of these relation combinations were added the candidate set. By following these steps, we extract 1,739 LVCs of the 1,768 LVC instances in the ON 4.99 data. Thus, we are detecting 98.42% of the LVCs in the gold annotations. Error analysis of the 28 instances that were not detected showed that they were missed either because of long-distance dependences and/or intervening relative clauses, and a few were annotation mistakes. The distribution of the three relation combinations is displayed in Table 3. The most frequent relation type is the direct object relation, which makes up 87.44% of all the LVCs.

5.2 Classifier

In LVC recognition, our classifier assigns a binary label (+1/-1) to an individual V_S-N_E pair on the dependency tree. Here, we use the machine learning algorithm called LibLinear, adopting L2-regularized logistic regression. This algorithm uses the following approach: given a set of instance-label pairs (x_i, y_i), where $x_i \in R^d$, $y_i \in$ {1,-1}, it finds a function $f : x_i \rightarrow y_i$ that maximizes the likelihood estimation of the classifier's parameters, which assumes that x_i was generated by a binomial model that depends on y_i (McCullagh and Nelder 1989). One advantage of using LibLinear over a Support Vector Machine (SVM) is the training and prediction speed. The dimensionality of our features is often very high, but the training sample size is small. LibLinear performs only logistic regression without using a kernel. Results show that LibLinear reduces both training and decoding times, while maintaining the accuracy of the prediction.

LEMMA: The lemma of V_S, N_E, V_S^{-1}, V_S^{+1}, N_E^{-1}, N_E^{+1}, V_S^h, N_E^h
POS: The part of speech tag of V_S, N_E, V_S^{-1}, V_S^{+1}, N_E^{-1}, N_E^{+1}, V_S^h, N_E^h
DEP: The dependents of V_S, N_E, V_S^{-1}, V_S^{+1}, N_E^{-1}, N_E^{+1}, V_S^p, N_E^h
RELPATH: The concatenation of the relation on the dependency tree path from V_S to N_E
POSPATH: The concatenation of the POS on the dependency tree path from V_S to N_E
DEPSUBCATSET: The subcategorization set that is derived by collecting all dependency labels of V_S and N_E
VOICE: Active or Passive for the V_S
DISTANCE: The dependency tree node distance between V_S and N_E

Table 4: Basic Features ($W^{-1/+1}$ refer to the left word / right word of W, and W^h refers to the head word of W).

Several features are used by the classifier, categorized into 3 different types: Basic Features, ON Word Sense Features, and WN Features.

Basic Features Basic features include the lexicon, part of speech (POS) tag, and the dependency relation of V_S and N_E. The paths of the dependency relation and POS are included as well. Additionally, the subcategorization frame which concatenates the dependency labels of V_S and N_E is adopted. These features are used either individually or jointly (e.g., POS of V_S and lemma of N_E make another new feature). The basic features are listed in Table 4.

OntoNotes Word Sense Features Word sense plays an important role in recognizing LVCs. For example, consider the following two sentences:

1. We are going to take a look at the trials and tribulations of Martha Stewart.
2. Barbie gets a makeover to give her a more youthful look.

In sentence (1) above, *take a look* is an LVC, while *give her a more youthful look* in sentence (2) is not. The difference in LVC status is reflected in the two different senses of *look*: the meaning of the first *look* is "act of looking," and the second usage of *look* is closer to the meaning "perceived appearance or feel." Although it is difficult to discover the lexical and dependency differences between these two V_S-N_E pairs, the word sense gives a useful clue for our classifier to identify the LVC.

In the ON 4.99 corpus, a word sense tag is annotated on the verbs with three or more senses and many nouns. The coarse-grained sense inventory, described in Section 3.3, gives a definition for each word sense. Ideally, for the data to be the most effective for LVC detection, all verbs and nouns would have sense tags. Unfortunately, not all of the subcorpora of the ON corpus are sense tagged. In the first step of our ON data experiment (in Section 6.2), the verbs and nouns in the automatically generated dependency trees don't contain any of the word sense tags. Hence, a Word Sense Disambiguation (WSD) model is necessary. In Lee (2002), a SVM-based WSD model that integrates the lemma, POS tag, and collocation information from near-by words is proposed. We apply this model to the WSD task with the ON word senses labels and implement our WSD classifier with the LibLinear algorithm with L2-regularization and L1-loss support vector classification. This algorithm uses a linear classifier to maximize the margin, instead of using a kernel. For the target word, we select ± 3 words as the window size, while we adopt the same feature list that was used in Lee (2002).

We train and test our model on the ON data for out-of-genre experiments (See Section 6 for the details of the data preparation). Our WSD model reaches a 76.16 Precision, 71.32 Recall, and 73.66 F_1 score. Although the overall performance of our WSD model is not ideal, the predicted word sense tag is only used in the automated generated dependency trees as one feature that supports the improvement of our LVC recognition.

WordNet Features WordNet contains rich word sense information and relational information between words. In our model, several pieces of WN information are used as features:

WordNet Sense: The fine-grained WN sense inventory provides word sense information for each verb and noun. As mentioned previously, the ON data is annotated with the ON sense inventory tag only. However, the ON sense inventory provides a mapping between each coarser-grained ON sense and the WN senses that it comprises. Thus, the WN sense tag can be extracted via the ON sense tag. Since the WN sense inventory is more fine-grained than the ON sense inventory, one ON sense may map to multiple WN senses. We opted to extract 1) The highest-frequency sense (with the lowest WN sense label number) as the WN Sense feature, and 2) The set of WN senses mapped to the ON sense. These two features are applied to both V_S and N_E.

WordNet Noun Type (Lexical File Information): For each of the noun senses in WN, the manually assigned lexical file information is given. These can be thought of as the word's supertype, and in this research, twelve basic types that indicate the noun could be eventive or stative are selected (discussed in Section 3.2). This base type is a more generalized property for each noun, and provides more common patterns for discovering previously unattested LVCs.

WordNet Hyponymy: Each word sense in WN contains the hypernym derived by the knowledge structure. The hypernym of the N_E provides a more generalized feature than the WN sense itself, but more fine-grained information than the base noun type.

6 Experiments and Results

For our experiments, we used two target corpora, the BNC LVC data provided by Tu and Roth (2011) and the ON 4.99 data. The BNC data is a balanced data set, including 1,039 positive LVC examples and 1,123 negative examples. We randomly sample 90% of the instances for training and the rest for testing. We also experiment with the ON 4.99 data. In order to evaluate the accuracy of our model for the different genres, we split our training and testing sets by randomly selecting different parts of subcorpora in each genre of ON. Portions of the following six corpora are used for the testing set: the MSNBC broadcast conversation, the CNN broadcast news, the Sinorama news magazine, the WSJ newswire, the CallHome telephone conversation, and the GALE web-text. In all of the ON data, 1,768 LVCs are annotated (in Table 3). Among all these LVCs in ON, 1,588 LVCs are listed in the training data set, and 180 LVCs are in the testing data set.

We also present an experiment investigating how to discover low-frequency LVCs using the WN synsets of nouns found in high-frequency LVCs (Section 6.3).

Model		P	R	F1
TR-C	+	**86.49**	84.21	85.33
	-	86.15	**88.19**	87.16
TR-S	+	86.48	85.09	86.46
	-	86.72	87.40	87.06
Basic	+	81.13	86.00	83.50
	-	88.89	84.85	86.82
All Features	+	85.32	**93.00**	**89.00**
	-	**94.31**	87.88	**90.98**

Table 5: Model Comparison for BNC data. *TR-C* is Tu & Roth's contextual feature model; *TR-S* refers is their statistical feature model. *Basic* model is our classifier with basic features only; compared to an *All Features* model. The '+' refers to performance in detecting LVCs, while the '-' refers to performance in detecting non-LVCs.

6.1 BNC Data

We first train and evaluate our model with the BNC data using automatic parsers produced by ClearNLP (Choi and Mccallum 2013). Table 5 shows the performance of Tu & Roth's model (2011) and our classifier on the BNC data set at each step: precision, recall, and F_1-measure. Our baseline model involves the basic features only. Our *All features* model, which includes the three WN features, gains around 3 to 4% improvement for positive and negative examples, with respect to Tu and Roth's contextual features and statistical features. In all we have added several beneficial features in comparison to the system of Tu & Roth: the dependency tree relation, the POS path between light verb and eventive noun, the subcategorization set, and the distance between light verb and eventive noun as new features. We discuss the the contribution of each individual feature in the next section.

6.2 OntoNotes Gold Data Evaluation

We first train and evaluate our model with automatic parse trees. The overall results are lower than on the BNC test set, in part due to errors in the automatic trees, but also because the data exhibits more variety with respect to the nouns found in the data, as discussed in Section 4. We achieved Precision of 54.94%, Recall of 77.22% and an F_1 score of

Feature	P	R	F1	Diff(%)
Basic	78.09	78.09	78.09	-
+ WN-Sense	80.23	**79.78**	80.00	+1.91
+ WN-Type	80.68	**79.78**	80.23	+0.23
+ WN-Hyper	81.61	**79.78**	**80.68**	+0.45
+ Word Sense	**81.77**	78.09	79.89	-0.79

Table 6: Incremental Feature Contribution for ON gold trees

Feature	P	R	F1	Diff(%)
Basic	78.09	78.09	78.09	-
WN-Sense	80.23	**79.78**	**80.00**	**+1.91**
WN-Type	78.53	78.09	78.31	+0.22
WN-Hyper	80.00	78.65	79.32	+1.23
Word Sense	**80.59**	76.97	78.74	+0.65

Table 7: Feature Contribution for ON gold trees

64.20%. We then use this data set with Gold Standard dependency trees to evaluate the contribution of the features individually. Table 6 shows the performance figures for our system with features added incrementally. These features are compared with the baseline model. All three WN features contribute to the F_1 score incrementally. After all the WN features are added, our model reaches the best F_1 score of 80.68. Although the addition of the ON Word Sense feature decreases the F_1 score, it still increases the precision.

To investigate the effectiveness of each individual feature, we carried out an ablation study using only one feature at a time. The results in Table 7 show that all the WN and ON word sense features improve the system's performance. This demonstrates that the more fine-grained features, including the WN-Sense, WN-Hyper, and ON Word Sense, contribute most to precision, especially the WN-Sense feature.

To understand the accuracy of our model on different light verb types, we conducted error analysis on the output for the best F_1 score in Table 6. As shown in Table 8, *make* and *have* achieve the highest and lowest F_1 scores, respectively. The poor results for *have* LVC recognition may be due to the borderline nature of many *have* LVCs, which makes them difficult to annotate consistently. For example, in *Why did we have a debate for a couple of days?*, it isn't clear whether *we* is a participant in the debate, or if they are simply holding the debate. Similarly, in *You could have a separate bilateral face to face negotiation with the United States...*, it is unclear whether *you* is a participant in the negotiation, or if this is a generic construction indicating the potential existence of negotiations. As shown previously in Table 2, *have* also combines with the greatest number of overlapping nouns that appear in both positive and negative LVC examples, making identification particularly difficult.

6.3 Using WordNet Relations

Although the above model could capture the majority of the LVCs in the corpus, those that are detected are relatively high-frequency LVCs. This led to the question of whether or not there is a better way to detect previously unattested LVCs. One idea is to leverage some general information that would allow the classifier to detect other possible LVCs. In the previous section, the results show that WN features provide positive contributions to our model. In this section, we analyze a small set of data from the ON corpus and corresponding WN features to explore the following possibility: if there is a high-frequency, attested *light verb* + *noun* combination, then any other eventive or stative noun sharing a synset or hypernym with this noun may also combine with that light verb to form an LVC.

Light Verb	P	R	F1
do	90.91	68.97	78.43
have	77.78	48.28	59.57
keep	100.00	50.00	66.67
make	79.55	97.22	87.50
take	82.22	82.22	82.22

Table 8: Error Analysis for Individual Light Verb Types

Data	# of V_S-N_E	# of LVC	Ratio (%)
LV + Synonym	91	49	53.85
LV + Noun	8,198	1,911	23.31

Table 9: LV + Synonym: number of potential LVC tokens detected by combining a particular light verb with all WN synonyms of a noun from a high-frequency LVC (# of V_S-N_E), compared with the number already annotated as LVCs (# of LVC). LV + Noun: number of potential LVC tokens detected by combining a light verb with any noun in ON (# of V_S-N_E), compared with the number already annotated as LVCs (# of LVC).

To explore this, we first calculate the frequency of all the gold LVC pairs in the ON 4.99 data. Then we extract the top 10 highest-frequency V_S-N_E pairs. In order to generate candidate LVC pairs, we fix the verb found in the high-frequency, attested LVC, and combine this with nouns that either share a synset or a hypernym with the noun from the same high-frequency LVC. This replacement of the eventive noun with its synonyms could allow for the discovery of promising LVC candidates. For example, the derived LVCs *make attempt*, *make effort*, and *make endeavor* are obtained by examining the synset and hypernym relations of the high-frequency LVC *make contribution*.

Using this process, we find a total of 91 tokens of potential LVCs in the ON corpus. When we compare this to our existing annotations, we see that 49 of these are already annotated as LVCs. Table 9 displays the numbers of gold true LVCs and candidate V_S-N_E pairs. The probability that combinations generated from the synonyms are true LVCs is twice the baseline probability that any V_S-N_E pair is an LVC. Thus, we can assume that WN synsets could play an important role in discovering low-frequency and previously unattested LVCs.

Notably, of the 91 potential LVC tokens that this process generated, there were 22 unique *verb* + *noun* types. Of these 22 potential LVC types, four were attested in the corpus and already annotated as LVCs, including the high-frequency types *make_effort* and *make_commitment*. This is not to say that the other candidate LVC combinations are not LVCs, but they are either not attested or not annotated as LVCs in the corpus. Further research is required.

7 Conclusion & Future Work

We have described a system for the recognition of LVCs in English, in which we build a regression classifier for automatically identifying LVCs based on lexical, WN, and ON word sense information. Our evaluations show that the performance of our system achieves an 88.90% F_1 score on the BNC data set and 64.20% F_1 score on the ON 4.99 data. Using ON Gold Standard parses and sense tags, our F_1 score is 80.68%. Evaluation also shows that both the WN and the ON word sense features result in better performance. In addition, we demonstrate that the LVCs derived by WN relations from high-frequency LVCs have a higher probability of true LVC-hood than other combinations of *light verb + noun*.

In the future, we would like to investigate adding more general information to our model, such as word embeddings and clustering based on verb dependencies. Secondly, we plan to integrate our LVC detection model into SRL processing to further improve the performance of the SRL system. We aim to improve our ON sense inventory and word sense disambiguation accuracy, and then apply it to our model. We will also update the ON 4.99 test set to include more consistent annotation of light usages of *give* and *get*.

Acknowledgments

We gratefully acknowledge the support of the National Science Foundation Grants IIS-1116782, A Bayesian Approach to Dynamic Lexical Resources for Flexible Language Processing, and 0910992 IIS:RI: Richer Representations for Machine Translation, and funding under the BOLT and Machine Reading programs, HR0011-11-C-0145 (BOLT) FA8750-09-C-0179 (M.R.). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

References

Choi, J. D., and Mccallum, A. 2013. Transition-based dependency parsing with selectional branching. In *In Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*.

Dligach, D., and Palmer, M. 2011. Good seed makes a good crop: Accelerating active learning using language modeling. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*.

Dowty, D. 1991. Thematic proto-roles and argument selection. *Language* 67:547–619.

Duffield, C.; Hwang, J.; Brown, S.; Viewig, S.; Davis, J.; and Palmer, M. 2007. Criteria for the manual grouping of verb senses. In *Proceedings of the Linguistic Annotation Workshop*.

Fan, R.-E.; Chang, K.-W.; Hsieh, C.-J.; Wang, X.-R.; and Lin, C.-J. 2008. LIBLINEAR: A library for large linear classification. *Journal of Machine Learning Research* 9:1871–1874.

Fellbaum, C.; Grabowski, J.; and Landes, S. 1998. Performance and confidence in a semantic annotation task. In Fellbaum, C., ed., *WordNet: An Electronic Database*. The MIT Press.

Gurrutxaga, A., and Alegria, I. 2012. Measuring the compositionality of nv expressions in basque by means of distributional similarity techniques. In *LREC*, 2389–2394.

Lee, Y. K., and Ng, H. T. 2002. An empirical evaluation of knowledge sources and learning algorithms for word sense disambiguation. In *Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing - Volume 10*, EMNLP '02, 41–48. Stroudsburg, PA, USA: Association for Computational Linguistics.

Marcus, M. P.; Santorini, B.; and Marcinkiewicz, M. A. 1994. Building a large annotated corpus of english: the penn treebank. *Computational Linguistics* 19.

McCullagh, P., and Nelder, J. A. 1989. *Generalized linear models (Second edition)*. London: Chapman & Hall.

Miller, G. A. 1995. Wordnet: a lexical database for english. *Communications of the ACM* 38(11):39–41.

Palmer, M.; Guildea, D.; and Kingsbury, P. 2005. The Proposition Bank: An annotated corpus of semantic roles. *Computational Linguistics* 31(1):71–105.

Pradhan, S.; Hovy, E.; Marcus, M.; M.Palmer; Ramshaw, L.; and Weschedel, R. 2007. Ontonotes: a unified relational semantic representation. In *Proceedings of the First IEEE International Conference on Semantic Computing(ICSC-07)*.

Tu, Y., and Roth, D. 2011. Learning english light verb constructions: Contextual or statistical. In *Proceedings of the Workshop on Multiword Expressions: from Parsing and Generation to the Real World*, 31–39. Association for Computational Linguistics.

Van de Cruys, T., and Moirón, B. V. 2007. Semantics-based multiword expression extraction. In *Proceedings of the Workshop on A Broader Perspective on Multiword Expressions*, 25–32. Association for Computational Linguistics.

Vincze, V.; Nagy T., I.; and Zsibrita, J. 2013. Learning to detect english and hungarian light verb constructions. *ACM Trans. Speech Lang. Process.* 10(2):6:1–6:25.

Weischedel, R.; Hovy, E.; Palmer, M.; Marcus, M.; Belvin, R.; Pradhan, S.; Ramshaw, L.; and Xue, N. 2011. Ontonotes: A large training corpus for enhanced processing. In Olive, J.; Christianson, C.; and McCary, J., eds., *Handbook of Natural Language Processing and Machine Translation*. Springer.