Learning English Light Verb Constructions: Contextual or Statistical

Yuancheng Tu
Department of Linguistics
University of Illinois
ytu@illinois.edu

Dan Roth
Department of Computer Science
University of Illinois
danr@illinois.edu

Abstract

In this paper, we investigate a supervised machine learning framework for automatically learning of English **L**ight **V**erb **C**onstructions (LVCs). Our system achieves an 86.3% accuracy with a baseline (chance) performance of 52.2% when trained with groups of either contextual or statistical features. In addition, we present an in-depth analysis of these contextual and statistical features and show that the system trained by these two types of cosmetically different features reaches similar performance empirically. However, in the situation where the surface structures of candidate LVCs are identical, the system trained with contextual features which contain information on surrounding words performs 16.7% better.

In this study, we also construct a balanced benchmark dataset with 2,162 sentences from BNC for English LVCs. And this data set is publicly available and is also a useful computational resource for research on MWEs in general.

1 Introduction

Multi-**W**ord **E**xpressions (MWEs) refer to various types of linguistic units or expressions, including idioms, noun compounds, named entities, complex verb phrases and any other habitual collocations. MWEs pose a particular challenge in empirical Natural Language Processing (NLP) because they always have idiosyncratic interpretations which cannot be formulated by directly aggregating the semantics of their constituents (Sag et al., 2002).

The study in this paper focuses on one special type of MWEs, i.e., the **L**ight **V**erb **C**onstructions (LVCs), formed from a commonly used verb and usually a noun phrase (NP) in its direct object position, such as *have a look* and *make an offer* in English. These complex verb predicates do not fall clearly into the discrete binary distinction of compositional or non-compositional expressions. Instead, they stand somewhat in between and are typically semi-compositional. For example, consider the following three candidate LVCs: *take a wallet*, *take a walk* and *take a while*. These three complex verb predicates are cosmetically very similar. But a closer look at their semantics reveals significant differences and each of them represents a different class of MWEs. The first expression, *take a wallet* is a literal combination of a verb and its object noun. The last expression *take a while* is an idiom and its meaning *cost a long time to do something*, cannot be derived by direct integration of the literal meaning of its components. Only the second expression, *take a walk* is an LVC whose meaning mainly derives from one of its components, namely its noun object (*walk*) while the meaning of its main verb is somewhat bleached (Butt, 2003; Kearns, 2002) and therefore *light* (Jespersen, 1965).

LVCs have already been identified as one of the major sources of problems in various NLP applications, such as automatic word alignment (Samardžić and Merlo, 2010) and semantic annotation transference (Burchardt et al., 2009), and machine translation. These problems provide empirical grounds for distinguishing between the bleached and full meaning of a verb within a given sentence, a task that is often difficult on the basis of surface structures since they always exhibit identical surface properties. For example, consider the following sentences:

Proceedings of the Workshop on Multiword Expressions: from Parsing and Generation to the Real World (MWE 2011), pages 31–39,
Portland, Oregon, USA, 23 June 2011. ©2011 Association for Computational Linguistics

1. He *had a look* of childish bewilderment on his face.
2. I've arranged for you to *have a look* at his file in our library.

In sentence 1, the verb *have* in the phrase *have a look* has its full fledged meaning "*possess, own*" and therefore it is *literal* instead of *light*. However, in sentence 2, *have a look* only means *look* and the meaning of the verb *have* is impoverished and is thus *light*.

In this paper, we propose an in-depth case study on LVC recognition, in which we investigate machine learning techniques for automatically identifying the impoverished meaning of a verb given a sentence. Unlike the earlier work that has viewed all verbs as possible light verbs (Tan et al., 2006), We focus on a half dozen of broadly documented and most frequently used English light verbs among the small set of them in English.

We construct a token-based data set with a total of $2,162$ sentences extracted from British National Corpus (BNC)[1] and build a learner with L2-loss SVM. Our system achieves a 86.3% accuracy with a baseline (chance) performance of 52.2%. We also extract automatically two groups of features, statistical and contextual features and present a detailed ablation analysis of the interaction of these features. Interestingly, the results show that the system performs similarly when trained independently with either groups of these features. And the integration of these two types of features does not improve the performance. However, when tested with all sentences with the candidate LVCs whose surface structures are identical in both negative and positive examples, for example, the aforementioned sentence 1 (negative) and 2 (positive) with the candidate LVC *"have a look"*, the system trained with contextual features which include information on surrounding words performs more robust and significantly better. This analysis contributes significantly to the understanding of the functionality of both contextual and statistical features and provides empirical evidence to guide the usage of them in NLP applications.

In the rest of the paper, we first present some related work on LVCs in Sec. 2. Then we describe our model including the learning algorithm and statistical and contextual features in Sec. 3. We present our experiments and analysis in Sec. 4 and conclude our paper in Sec. 5.

2 Related Work

LVCs have been well-studied in linguistics since early days (Jespersen, 1965; Butt, 2003; Kearns, 2002). Recent computational research on LVCs mainly focuses on type-based classification, i.e., statistically aggregated properties of LVCs. For example, many works are about direct measuring of the compositionality (Venkatapathy and Joshi, 2005), compatibility (Barrett and Davis, 2003), acceptability (North, 2005) and productivity (Stevenson et al., 2004) of LVCs. Other works, if related to token-based identification, i.e., identifying idiomatic expressions within context, only consider LVCs as one small subtype of other idiomatic expressions (Cook et al., 2007; Fazly and Stevenson, 2006).

Previous computational works on token-based identification differs from our work in one key aspect. Our work builds a learning system which systematically incorporates both informative statistical measures and specific local contexts and does in-depth analysis on both of them while many previous works, either totally rely on or only emphasize on one of them. For example, the method used in (Katz and Giesbrecht, 2006) relies primarily on local co-occurrence lexicon to construct feature vectors for each target token. On the other hand, some other works (Fazly and Stevenson, 2007; Fazly and Stevenson, 2006; Stevenson et al., 2004), argue that linguistic properties, such as canonical syntactic patterns of specific types of idioms, are more informative than local context.

Tan et.al. (Tan et al., 2006) propose a learning approach to identify token-based LVCs. The method is only similar to ours in that it is a supervised framework. Our model uses a different data set annotated from BNC and the data set is larger and more balanced compared to the previous data set from WSJ. In addition, previous work assumes all verbs as potential LVCs while we intentionally exclude those verbs which linguistically never tested as light verbs, such as *buy* and *sell* in English and only focus on a half dozen of broadly documented English light

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

verbs, such as *have*, *take*, *give*, *do*, *get* and *make*.

The lack of common benchmark data sets for evaluation in MWE research unfortunately makes many works incomparable with the earlier ones. The data set we construct in this study hopefully can serve as a common test bed for research in LVCs or MWEs in general.

3 Learning English LVCs

In this study, we formulate the context sensitive English LVC identification task as a supervised binary classification problem. For each target LVC candidate within a sentence, the classifier decides if it is a true LVC. Formally, given a set of n labeled examples $\{x_i, y_i\}_{i=1}^{n}$, we learn a function $f : \mathcal{X} \rightarrow \mathcal{Y}$ where $\mathcal{Y} \in \{-1, 1\}$. The learning algorithm we use is the classic soft-margin SVM with L2-loss which is among the best "off-the-shelf" supervised learning algorithms and in our experiments the algorithm indeed gives us the best performance with the shortest training time. The algorithm is implemented using a modeling language called Learning Based Java (LBJ) (Rizzolo and Roth, 2010) via the LIBSVM Java API (Chang and Lin, 2001).

Previous research has suggested that both local contextual and statistical measures are informative in determining the class of an MWE token. However, it is not clear to what degree these two types of information overlap or interact. Do they contain similar knowledge or the knowledge they provide for LVC learning is different? Formulating a classification framework for identification enables us to integrate all contextual and statistical measures easily through features and test their effectiveness and interaction systematically.

We focus on two types of features: contextual and statistical features, and analyze in-depth their interaction and effectiveness within the learning framework. Statistical features in this study are numerical features which are computed globally via other big corpora rather than the training and testing data used in the system. For example, the *Cpmi* and *Deverbal v/n Ratio* (details in sec. 3.1) are generated from the statistics of Google n-gram and BNC corpus respectively. Since the *phrase size* feature is numerical and the selection of the candidate LVCs in the data set uses the canonical length information[2], we include it into the statistical category. Contextual features are defined in a broader sense and consist of all local features which are generated directly from the input sentences, such as word features within or around the candidate phrases. We describe the details of the used contextual features in sec. 3.2.

Our experiments show that arbitrarily combining statistic features within our current learning system does not improve the performance. Instead, we provide systematic analysis for these features and explore some interesting empirical observations about them within our learning framework.

3.1 Statistical Features

Cpmi: *C*ollocational *p*oint-wise *m*utual *i*nformation is calculated from Google n-gram dataset whose n-gram counts are generated from approximately one trillion words of text from publicly accessible Web pages. We use this big data set to overcome the data sparseness problem.

Previous works (Stevenson et al., 2004; Cook et al., 2007) show that one canonical surface syntactic structure for LVCs is *V + a/an Noun*. For example, in the LVC *take a walk*, "take" is the verb (V) and "walk" is the deverbal noun. The typical determiner in between is the indefinite article "a". It is also observed that when the indefinite article changes to definite, such as "the", "this" or "that", a phrase is less acceptable to be a true LVC. Therefore, the direct collocational pmi between the verb and the noun is derived to incorporate this intuition as shown in the following[3]:

$$Cpmi = 2I(v, aN) - I(v, theN)$$

Within this formula, $I(v, aN)$ is the point-wise mutual information between "v", the verb, and "aN", the phrase such as "a walk" in the aforementioned example. Similar definition applies to $I(v, theN)$. PMI of a pair of elements is calculated as (Church et al., 1991):

$$I(x, y) = \log \frac{N_{x+y} f(x, y)}{f(x, *) f(*, y)}$$

[2] We set an empirical length constraint to the maximal length of the noun phrase object when generating the candidates from BNC corpus.

[3] The formula is directly from (Stevenson et al., 2004).

N_{x+y} is the total number of verb and a/the noun pairs in the corpus. In our case, all trigram counts with this pattern in N-gram data set. $f(x, y)$ is the frequency of x and y co-occurring as a v-a/theN pair where $f(x, *)$ and $f(*, y)$ are the frequency when either of x and y occurs independent of each other in the corpus. Notice these counts are not easily available directly from search engines since many search engines treat articles such as "a" or "the" as stop words and remove them from the search query[4].

Deverbal v/n Ratio: the second statistical feature we use is related to the verb and noun usage ratio of the noun object within a candidate LVC. The intuition here is that the noun object of a candidate LVC has a strong tendency to be used as a verb or related to a verb via derivational morphology. For example, in the candidate phrase "have a look", "look" can directly be used as a verb while in the phrase "make a transmission", "transmission" is derivationally related to the verb "transmit". We use frequency counts gathered from British National Corpus (BNC) and then calculate the ratio since BNC encodes the lexeme for each word and is also tagged with parts of speech. In addition, it is a large corpus with 100 million words, thus, an ideal corpus to calculate the verb-noun usage for each candidate word in the object position.

Two other lexical resources, WordNet (Fellbaum, 1998) and NomLex (Meyers et al., 1998), are used to identify words which can directly be used as a noun and a verb and those that are derivational related. Specifically, WordNet is used to identify the words which can be used as both a noun and a verb and NomLex is used to recognize those derivationally related words. And the verb usage counts of these nouns are the frequencies of their corresponding derivational verbs. For example, for the word "transmission", its verb usage frequency is the count in BNC with its derivationally related verb "transmit".

Phrase Size: the third statistical feature is the actual size of the candidate LVC phrase. Many modifiers can be inserted inside the candidate phrases to generate new candidates. For example, "take a look" can be expanded to "take a *close* look", "take an *extremely* close look" and the expansion is in theory infinite. The hypothesis behind this feature is that regular usage of LVCs tends to be short. For example, it is observed that the canonical length in English is from 2 to 6.

3.2 Contextual Features

All features generated directly from the input sentences are categorized into this group. They consists of features derived directly from the candidate phrases themselves as well as their surrounding contexts.

Noun Object: this is the noun head of the object noun phrase within the candidate LVC phrase. For example, for a verb phrase "take a quick look", its noun head "look" is the active *Noun Object* feature. In our data set, there are 777 distinctive such nouns.

LV-NounObj: this is the bigram of the light verb and the head of the noun phrase. This feature encodes the collocation information between the candidate light verb and the head noun of its object.

Levin's Class: it is observed that members within certain groups of verb classes are legitimate candidates to form acceptable LVCs (Fazly et al., 2005). For example, many sound emission verbs according to Levin (Levin, 1993), such as *clap*, *whistle*, and *plop*, can be used to generate legitimate LVCs. Phrases such as *make a clap/plop/whistle* are all highly acceptable LVCs by humans even though some of them, such as *make a plop* rarely occur within corpora. We formulate a vector for all the 256 Levin's verb classes and turn the corresponding class-bits on when the verb usage of the head noun in a candidate LVC belongs to these classes. We add one extra class, *other*, to be mapped to those verbs which are not included in any one of these 256 Levin's verb classes.

Other Features: we construct other local contextual features, for example, the part of speech of the word immediately before the light verb (titled *posBefore*) and after the whole phrase (*posAfter*). We also encode the determiner within all candidate LVCs as another lexical feature (*Determiner*). We examine many other combinations of these contextual features. However, only those features that contribute positively to achieve the highest performance of the classifier are listed for detailed analysis in the next section.

[4] Some search engines accept "quotation strategy" to retain stop words in the query.

4 Experiments and Analysis

In this section, we report in detail our experimental settings and provide in-depth analysis on the interactions among features. First, we present our motivation and methodology to generate the new data set. Then we describe our experimental results and analysis.

4.1 Data Preparation and Annotation

The data set is generated from BNC, a balanced synchronic corpus containing 100 million words collected from various sources of British English. We begin our sentence selection process with the examination of a handful of previously investigated verbs (Fazly and Stevenson, 2007; Butt, 2003). Among them, we pick the 6 most frequently used English light verbs: *do*, *get*, *give*, *have*, *make* and *take*.

To identify potential LVCs within sentences, we first extract all sentences where one or more of the six verbs occur from BNC (XML Edition) and then parse these sentences with Charniak's parser (Charniak and Johnson, 2005). We focus on the "*verb + noun object*" pattern and choose all the sentences which have a direct NP object for the target verbs. We then collect a total of 207,789 sentences.

We observe that within all these chosen sentences, the distribution of true LVCs is still low. We therefore use three resources to filter out trivial negative examples. Firstly, We use WordNet (Fellbaum, 1998) to identify the head noun in the object position which can be used as both a noun and a verb. Then, we use frequency counts gathered from BNC to filter out candidates whose verb usage is smaller than their noun usage. Finally, we use NomLex (Meyers et al., 1998) to recognize those head words in the object position whose noun forms and verb forms are derivationally related, such as *transmission* and *transmit*. We keep all candidates whose object head nouns are derivationlly related to a verb according to a gold-standard word list we extract from NomLex[5]. With this pipeline method, we filter out approximately 55% potential negative examples. This leaves us with 92,415 sentences which we sample about 4% randomly to present to annotators. This filtering method successfully improves the recall of the positive examples and ensures us a corpus with balanced examples.

A website[6] is set up for annotators to annotate the data. Each potential LVC is presented to the annotator in a sentence. The annotator is asked to decide whether this phrase within the given sentence is an LVC and to choose an answer from one of these four options: *Yes*, *No*, *Not Sure*, and *Idiom*.

Detailed annotation instructions and LVC examples are given on the annotation website. When facing difficult examples, the annotators are instructed to follow a general "*replacing*" principle, i.e, if the candidate light verb within the sentence can be replaced by the verb usage of its direct object noun and the meaning of the sentence does not change, that verb is regarded as a light verb and the candidate is an LVC. Each example is annotated by two annotators and We only accept examples where both annotators agree on positive or negative. We generate a total of 1,039 positive examples and 1,123 negative examples. Among all these positive examples, there are 760 distinctive LVC phrases and 911 distinctive verb phrases with the pattern *"verb + noun object"* among negative examples. The generated data set therefore gives the classifier the 52.2% chance baseline if the classifier always votes the majority class in the data set.

4.2 Evaluation Metrics

For each experiment, we evaluate the performance with three sets of metrics. We first report the standard accuracy on the test data set. Since accuracy is argued not to be a sufficient measure of the evaluation of a binary classifier (Fazly et al., 2009) and some previous works also report F1 values for the positive classes, we therefore choose to report the precision, recall and F1 value for both positive and negative classes.

		True Class	
		+	-
Predicted Class	+	**tp**	**fp**
	-	**fn**	**tn**

Table 1: Confusion matrix to define *true positive (tp)*, *true negative (tn)*, *false positive (fp)* and *false negative (fn)*.

[5] We do not count those nouns ending with *er* and *ist*

[6] http://cogcomp.cs.illinois.edu/∼ytu/test/LVCmain.html

Based on the classic confusion matrix as shown in Table 1, we calculate the precision and recall for the positive class in equation 1:

$$P^{+} = \frac{tp}{tp + fp} \qquad R^{+} = \frac{tp}{tp + fn} \tag{1}$$

And similarly, we use equation 2 for negative class. And the F1 value is the harmonic mean of the precision and recall of each class.

$$P^{-} = \frac{tn}{tn + fn} \qquad R^{-} = \frac{tn}{tn + fp} \tag{2}$$

4.3 Experiments with Contextual Features

In our experiments, We aim to build a high performance LVC classifier as well as to analyze the interaction between contextual and statistical features. We randomly sample 90% sentences for training and the rest for testing. Our chance baseline is 52.2%, which is the percentage of our majority class in the data set. As shown in Table 2, the classifier reaches an 86.3% accuracy using all contextual features described in previous section 3.2. Interestingly, we observe that adding other statistical features actually hurts the performance. The classifier can effectively learn when trained with discrete contextual features.

Label	Precision	Recall	F1
+	86.486	84.211	85.333
-	86.154	88.189	87.160
Accuracy	**86.307**		
Chance Baseline	52.2		

Table 2: By using all our contextual features, our classifier achieves overall 86.307% accuracy.

In order to examine the effectiveness of each individual feature, we conduct an ablation analysis and experiment to use only one of them each time. It is shown in Table 3 that *LV-NounObj* is found to be the most effective contextural feature since it boosts the baseline system up the most, an significant increase of 31.6%.

We then start from this most effective feature, *LV-NounObj* and add one feature each step to observe the change of the system accuracy. The results are listed in Table 4. Other significant features are features within the candidate LVCs themselves such as *Determiner, Noun Object* and *Levin's Class* related to the object noun. This observation agrees with previous research that the acceptance of LVCs is closely correlated to the linguistic properties of their components. The part of speech of the word after the phrase seems to have negative effect on the performance. However, experiments show that without this feature, the overall performance decreases.

Features	Accuracy	Diff(%)
Baseline (chance)	52.2	
LV-NounObj	83.817	**+31.6**
Noun Object	79.253	+27.1
Determiner	72.614	+20.4
Levin's Class	69.295	+17.1
posBefore	53.112	+0.9
posAfter	51.037	-1.1

Table 3: Using only one feature each time. *LV-NounObj* is the most effective feature. Performance gain is associated with a plus sign and otherwise a negative sign.

Features	Accuracy	Diff(%)
Baseline (chance)	52.2	
+ LV-NounObj	83.817	+*31.6*
+ Noun Object	84.232	+*0.4*
+ Levin's Class	84.647	+*0.4*
+ posBefore	84.647	0.0
+ posAfter	83.817	-0.8
+ Determiner	86.307	+*2.5*

Table 4: Ablation analysis for contextual features. Each feature is added incrementally at each step. Performance gain is associated with a plus sign otherwise a negative sign.

4.4 Experiments with Statistical Features

When using statistical features, instead of directly using the value, we discretize each value to a binary feature. On the one hand, our experiments show that this way of transformation achieves the best performance. On the other hand, the transformation plays an analogical role as a kernel function which maps one dimensional non-linear separable examples into an infinite or high dimensional space to render the data linearly separable.

In these experiments, we use only numerical features described in section 3.1. And it is interesting to observe that those features achieve very similar

Label	Precision	Recall	F1
+	86.481	85.088	86.463
-	86.719	87.402	87.059
Accuracy	**86.307**		

Table 5: Best performance achieved with statistical features. Comparing to Table 2, the performance is similar to that trained with all contextual features.

performance as the contextual features as shown in Table 5.

To validate that the similar performance is not incidental. We then separate our data into 10-fold training and testing sets and learn independently from each fold of these ten split. Figure 1, which shows the comparison of accuracies for each data fold, indicates the comparable results for each fold of the data. Therefore, we conclude that the similar effect achieved by training with these two groups of features is not accidental.

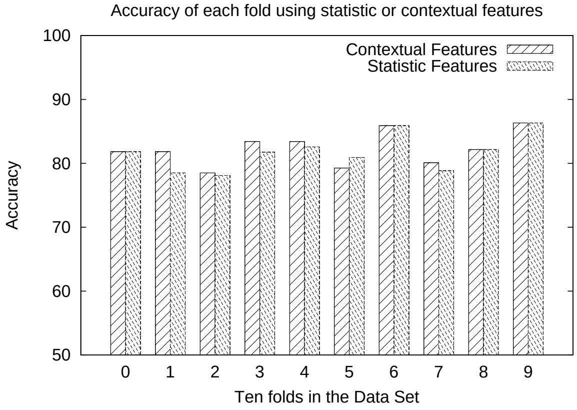

Figure 1: Classifier Accuracy of each fold of all 10 fold testing data, trained with groups of statistical features and contextual features separately. The similar height of each histogram indicates the similar performance over each data separation and the similarity is not incidental.

We also conduct an ablation analysis with statistical features. Similar to the ablation analyses for contextual features, we first find that the most effective statistical feature is *Cpmi*, the collocational based point-wise mutual information. Then we add one feature at each step and show the increasing performance in Table 6. *Cpmi* is shown to be a good indicator for LVCs and this observation agrees with many previous works on the effectiveness of point-wise mutual information in MWE identification tasks.

Features BaseLine (chance)	Accuracy 52.2	Diff(%)
+ Cpmi	83.402	+**31.2**
+ Deverbal v/n Ratio	85.892	+2.5
+ Phrase Size	86.307	+0.4

Table 6: Ablation analysis for statistical features. Each feature is added incrementally at each step. Performance gain is associated with a plus sign.

4.5 Interaction between Contextual and Statistical Features

Experiments from our previous sections show that two types of features which are cosmetically different actually achieve similar performance. In the experiments described in this section, we intend to do further analysis to identify further the relations between them.

4.5.1 Situation when they are similar

Our ablation analysis shows that *Cpmi* and *LV-NounObj* features are the most two effective features since they boost the baseline performance up more than 30%. We then train the classifier with them together and observe that the classifier exhibits similar performance as the one trained with them independently as shown in Table 7. This result indicates that these two types of features actually provide similar knowledge to the system and therefore combining them together does not provide any additional new information. This observation also agrees with the intuition that point-wise mutual information basically provides information on word collocations (Church and Hanks, 1990).

Feature	Accuracy	F1+	F1-
LV-NounObj	83.817	82.028	85.283
Cpmi	83.402	81.481	84.962
Cpmi+LV-NounObj	83.817	82.028	85.283

Table 7: The classifier achieves similar performance trained jointly with *Cpmi* and *LV-NounObj* features, comparing with the performance trained independently.

4.5.2 Situation when they are different

Token-based LVC identification is a difficult task on the basis of surface structures since they always exhibit identical surface properties. However, candidate LVCs with identical surface structures in both positive and negative examples provide an ideal test bed for the functionality of local contextual features. For example, consider again these two aforementioned sentences which are repeated here for reference:

1. He *had a look* of childish bewilderment on his face.
2. I've arranged for you to *have a look* at his file in our library.

The system trained only with statistic features cannot distinguish these two examples since their type-based statistical features are exactly the same. However, the classifier trained with local contextual features is expected to perform better since it contains feature information from surrounding words. To verify our hypothesis, we extract all examples in our data set which have this property and then select same number of positive and negative examples from them to formulate our test set. We then train out classifier with the rest of the data, independently with contextual features and statistical features. As shown in Table 8, the experiment results validate our hypothesis and show that the classifier trained with contextual features performs significantly better than the one trained with statistical features. The overall lower system results also indicate that indeed the test set with all ambiguous examples is a much harder test set.

One final observation is the extremely low F1 value for negative class and relatively good performance for positive class when trained with only statistical features. This may be explained by the fact that statistical features have stronger bias toward predicting examples as positive and can be used as an unsupervised metric to acquire real LVCs in corpora.

Classifier	Accuracy	F1+	F1-
Contextual	**68.519**	75.362	56.410
Statistical	51.852	88.976	27.778
Diff (%)	+16.7	-13.6	+28.3

Table 8: Classifier trained with local contextual features is more robust and significantly better than the one trained with statistical features when the test data set consists of all ambiguous examples.

5 Conclusion and Further Research

In this paper, we propose an in-depth case study on LVC recognition, in which we build a supervised learning system for automatically identifying LVCs in context. Our learning system achieves an 86.3% accuracy with a baseline (chance) performance of 52.2% when trained with groups of either contextual or statistical features. In addition, we exploit in detail the interaction of these two groups of contextual and statistical features and show that the system trained with these two types of cosmetically different features actually reaches similar performance in our learning framework. However, when it comes to the situation where the surface structures of candidate LVCs are identical, the system trained with contextual features which include information on surrounding words provides better and more robust performance.

In this study, we also construct a balanced benchmark dataset with 2,162 sentences from BNC for token-based classification of English LVCs. And this data set is publicly available and is also a useful computational resource for research on MWEs in general.

There are many aspects for further research of the current study. One direction for further improvement would be to include more long-distance features, such as parse tree path, to test the sensitivity of the LVC classifier to those features and to examine more extensively the combination of the contextual and statistical features. Another direction would be to adapt our system to other MWE types and to test if the analysis on contextual and statistical features in this study also applies to other MWEs.

Acknowledgments

The authors would like to thank all annotators who annotated the data via the web interface and four annonymous reviewers for their valuable comments. The research in this paper was supported by the Multimodal Information Access & Synthesis Center at

UIUC, part of CCICADA, a DHS Science and Technology Center of Excellence.

References

L. Barrett and A. Davis. 2003. Diagnostics for determing compatibility in english support verb nominalization pairs. In *Proceedings of CICLing-2003*, pages 85–90.

A. Burchardt, K. Erk, A. Frank, A. Kowalski, S. Pado, and M. Pinkal. 2009. Using framenet for semantic analysis of german: annotation, representation and automation. In Hans Boas, editor, *Multilingual FrameNets in Computational Lexicography: methods and applications*, pages 209–244. Mouton de Gruyter.

M. Butt. 2003. The light verb jungle. In *Harvard Working Paper in Linguistics*, volume 9, pages 1–49.

C. Chang and C. Lin, 2001. *LIBSVM: a library for support vector machines*. Software available at http://www.csie.ntu.edu.tw/∼cjlin/libsvm.

E. Charniak and M. Johnson. 2005. Coarse-to-fine n-best parsing and maxent discriminative reranking. In *Proceedings of ACL-2005*.

K. Church and P. Hanks. 1990. Word association norms, mutual information, and lexicography. *Computational Linguistics*, 16(1), March.

K. Church, W. Gale, P. Hanks, and D. Hindle. 1991. Using statistics in lexical analysis. In *Lexical Acquisition: Exploiting On-Line Resources to Build a Lexicon*, pages 115–164. Erlbaum.

P. Cook, A. Fazly, and S. Stevenson. 2007. Pulling their weight: Exploiting syntactic forms for the automatic identification of idiomatic expressions in context. In *Proceedings of the Workshop on A Broader Perspective on Multiword Expressions*, pages 41–48, Prague, Czech Republic, June. Association for Computational Linguistics.

A. Fazly and S. Stevenson. 2006. Automatically constructing a lexicon of verb phrase idiomatic combinations. In *Proceedings of EACL-2006*.

A. Fazly and S. Stevenson. 2007. Distinguishing subtypes of multiword expressions using linguistically-motivated statistical measures. In *Proceedings of the Workshop on A Broader Perspective on Multiword Expressions*, pages 9–16, Prague, Czech Republic, June.

A. Fazly, R. North, and S. Stevenson. 2005. Automatically distinguishing literal and figurative usages of highly polysemous verbs. In *Proceedings of the ACL-SIGLEX Workshop on Deep Lexical Acquisition*, pages 38–47, Ann Arbor, Michigan, June. Association for Computational Linguistics.

A. Fazly, P. Cook, and S. Stevenson. 2009. Unsupervised type and token identification of idiomatic expression. *Comutational Linguistics*.

C. Fellbaum, editor. 1998. *WordNet: An Electronic Lexical Database*. MIT Press.

O. Jespersen. 1965. *A Modern English Grammar on Historical Principles, Part VI, Morphology*. Aeorge Allen and Unwin Ltd.

G. Katz and E. Giesbrecht. 2006. Automatic identification of non-compositional multi-word expressions using latent semantic analysis. In *Proceedings of the Workshop on Multiword Expressions: Identifying and Exploiting Underlying Properties*, pages 12–19.

K. Kearns. 2002. Light verbs in english. In *http://www.ling.canterbury.ac.nz/documents*.

B. Levin. 1993. *English Verb Classes and Alternations, A Preliminary Investigation*. University of Chicago Press.

A. Meyers, C. Macleod, R. Yangarber, R. Grishman, L. Barrett, and R. Reeves. 1998. Using nomlex to produce nominalization patterns for information extraction. In *Proceedings of COLING-ACL98 Workshop:the Computational Treatment of Nominals*.

R. North. 2005. Computational measures of the acceptability of light verb constructions. University of Toronto, Master Thesis.

N. Rizzolo and D. Roth. 2010. Learning based java for rapid development of nlp systems. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC)*.

I. Sag, T. Baldwin, F. Bond, and A. Copestake. 2002. Multiword expressions: A pain in the neck for nlp. In *In Proc. of the 3rd International Conference on Intelligent Text Processing and Computational Linguistics (CICLing-2002*, pages 1–15.

T. Samardžić and P. Merlo. 2010. Cross-lingual variation of light verb constructions: Using parallel corpora and automatic alignment for linguistic research. In *Proceedings of the 2010 Workshop on NLP and Linguistics: Finding the Common Ground*, pages 52–60, Uppsala, Sweden, July.

S. Stevenson, A. Fazly, and R. North. 2004. Statistical measures of the semi-productivity of light verb constructions. In *Proceedings of ACL-04 workshop on Multiword Expressions: Integrating Processing*, pages 1–8.

Y. Tan, M. Kan, and H. Cui. 2006. Extending corpus-based identification of light verb constructions using a supervised learning framework. In *Proceedings of EACL-06 workshop on Multi-word-expressions in a multilingual context*, pages 49–56.

S. Venkatapathy and A. Joshi. 2005. Measuring the relative compositionality of verb-noun (v-n) collocations by integrating features. In *Proceedings of HLT and EMNLP05*, pages 899–906.