Extending corpus-based identification of light verb constructions using a supervised learning framework

Yee Fan Tan, **Min-Yen Kan** and **Hang Cui**
Department of Computer Science
National University of Singapore
3 Science Drive 2, Singapore 117543
{tanyeefa, kanmy, cuihang}@comp.nus.edu.sg

Abstract

Light verb constructions (LVCs), such as "make a call" in English, can be said to be complex predicates in which the verb plays only a functional role. LVCs pose challenges for natural language understanding, as their semantics differ from usual predicate structures. We extend the existing corpus-based measures for identifying LVCs between verb-object pairs in English, by proposing using new features that use mutual information and assess other syntactic properties. Our work also incorporates both existing and new LVC features into a machine learning approach. We experimentally show that using the proposed framework incorporating all features outperforms previous work by 17%. As machine learning techniques model the trends found in training data, we believe the proposed LVC detection framework and statistical features is easily extendable to other languages.

1 Introduction

Many applications in natural language processing rely on the relationships between words in a document. Verbs play a central role in many such tasks; for example, the assignment of semantic roles to noun phrases in a sentence heavily depends on the verb that link the noun phrases together (as in "Pierre Vinken/SUBJ, will join/PRED, the board/OBJ").

However, verb processing is difficult because of many phenomena, such as normalization of actions, verb particle constructions and light verb constructions. Applications that process verbs must handle these cases effectively. We focus on the identification of light verb constructions (also known as support verb constructions) in English, as such constructions play a prominent and productive role in many other languages (Butt and Geuder, 2001; Miyamoto, 2000). Although the exact definition of a LVC varies in the literature, we use the following operational definition:

> A **light verb construction (LVC)** is a verb-complement pair in which the verb has little lexical meaning (is "light") and much of the semantic content of the construction is obtained from the complement.

Examples of LVCs in English include "give a speech", "make good (on)" and "take (NP) into account". In the case in which the complement is a noun, it is often a deverbal noun and, as such, can usually be paraphrased using the object's root verb form without (much) loss in its meaning (*e.g.*, take a walk → walk, make a decision → decide, give a speech → speak).

We propose a corpus-based approach to determine whether a verb-object pair is a LVC. Note that we limit the scope of LVC detection to LVCs consisting of verbs with noun complements. Specifically, we extend previous work done by others by examining how the local context of the candidate construction and the corpus-wide frequency of related words to the construction play an influence on the lightness of the verb.

A second contribution is to integrate our new features with previously reported ones under a machine learning framework. This framework optimizes the weights for these measures automatically against a training corpus in supervised learning, and attests to the significant modeling improvements of our features on our corpus. Our corpus-based evaluation shows that the combination of previous work and our new features improves LVC detection significantly over previous work.

An advantage gained by adopting a machine learning framework is that it can be easily adapted to other languages that also exhibit light verbs. While we perform evaluations on English, light verbs exist in most other languages. In some of these languages, such as Persian, most actions are expressed as LVCs rather than single-word verbs (Butt, 2003). As such, there is currently a unmet demand for developing an adaptable framework for LVC detection that applies across languages. We believe the features proposed in this paper would also be effective in identifying light verbs in other languages.

We first review previous corpus-based approaches to LVC detection in Section 2. In Section 3, we show how we extend the use of mutual information and employ context modeling as features for improved LVC detection. We next describe our corpus processing and how we compiled our gold standard judgments used for supervised machine learning. In Section 4, we evaluate several feature combinations before concluding the paper.

2 Related Work

With the recent availability of large corpora, statistical methods that leverage syntactic features are a current trend. This is the case for LVC detection as well.

Grefenstette and Teufel (1995) considered a similar task of identifying the most probable light verb for a given deverbal noun. Their approach focused on the deverbal noun and occurrences of the noun's verbal form, arguing that the deverbal noun retains much of the verbal characteristics in the LVCs. To distinguish the LVC from other verb-object pairs, the deverbal noun must share similar argument/adjunct structures with its verbal counterpart. Verbs that appear often with these characteristic deverbal noun forms are deemed light verbs. They approximate the identification of argument/adjunct structures by using the preposition head of prepositional phrases that occur after the verb or object of interest.

Let n be a deverbal noun whose most likely light verb is to be found. Denote its verbal form by v', and let P be the set containing the three most frequently occurring prepositions that occur after v'. The verb-object pairs that are not followed by a preposition in P are filtered out. For any verb v, let $g(v,n)$ be the count of verb-object pairs v-n that remain after the filtering step above. Grefenstette and Teufel proposed that the light verb for n be returned by the following equation:

$$\text{GT95}(n) = \arg\max_{v} g(v,n) \tag{1}$$

Interestingly, Grefenstette and Teufel indicated that their subsequent experiments suggested that the filtering step may not be necessary.

Whereas the GT95 measure centers on the deverbal object, Dras and Johnson (1996) also consider the verb's corpus frequency. The use of this complementary information improves LVC identification, as it models the inherent bias of some verbs to be used more often as light verbs than others. Let $f(v,n)$ be the count of verb-object pairs occurring in the corpus, such that v is the verb, n is a deverbal noun. Then, the most probable light verb for n is given by:

$$\text{DJ96}(n) = \arg\max_{v} f(v,n) \sum_{n} f(v,n) \tag{2}$$

Stevenson *et al.* (2004)'s research examines evidence from constructions featuring determiners. They focused on expressions of the form v-a-n and v-det-n, where v is a light verb, n is a deverbal noun, a is an indefinite determiner (namely, "a" or "an"), and det is any determiner other than the indefinite. Examples of such constructions are "give a speech" and "take a walk". They employ mutual information which measures the frequency of co-occurrences of two variables, corrected for random agreement. Let $I(x,y)$ be the mutual information between x and y. Then the following measure can be used:

$$\text{SFN04}(v,n) = 2 \times I(v, a\text{-}n) - I(v, det\text{-}n), \tag{3}$$

where higher values indicate a higher likelihood of v-a-n being a light verb construction. Also, they suggested that the determiner "the" be excluded from the development data since it frequently occurred in their data.

Recently, Fazly *et al.* (2005) have proposed a statistical measure for the detection of LVCs. The probability that a verb-object pair v-n (where v is a light verb) is a LVC can be expressed as a product of three probabilities: (1) probability of the object

n occurring in the corpus, (2) the probability that n is part of any LVC given n, and (3) the probability of v occurring given n and that v-n is a LVC. Each of these three probabilities can then be estimated by the frequency of occurrence in the corpus, using the assumption that all instances of v'-a-n is a LVC, where v' is any light verb and a is an indefinite determiner.

To summarize, research in LVC detection started by developing single measures that utilized simple frequency counts of verbs and their complements. From this starting point, research has developed in two different directions: using more informed measures for word association (specifically, mutual information) and modeling the context of the verb-complement pair.

Both the GT95 and DJ96 measures suffer from using frequency counts directly. Verbs that are not light but occur very frequently (such as "buy" and "sell" in the Wall Street Journal) will be marked by these measures. As such, given a deverbal noun, they sometimes suggest verbs that are not light. We hypothesize that substituting MI for frequency count can alleviate this problem.

The SFN04 metric adds in the context provided by determiners to augment LVC detection. This measure may work well for LVCs that are marked by determiners, but excludes a large portion of LVCs that are composed without determiners. To design a robust LVC detector requires integrating such specific contextual evidence with other general evidence.

Building on this, Fazly *et al.* (2005) incorporate an estimation of the probability that a certain noun is part of a LVC. However, like SFN04, LVCs without determiners are excluded.

3 Framework and Features

Previous work has shown that different measures based on corpus statistics can assist in LVC detection. However, it is not clear to what degree these different measures overlap and can be used to reinforce each other's results. We solve this problem by viewing LVC detection as a supervised classification problem. Such a framework can integrate the various measures and enable us to test their combinations in a generic manner. Specifically, each verb-object pair constitutes an individual classification instance, which possesses a set of features $f_1, \ldots, f_n$ and is assigned a class label from the binary classification of $\{LVC, \neg LVC\}$.

In such a machine learning framework, each of the aforementioned metrics are separate features.

In our work, we have examined three different sets of features for LVC classification: (1) base, (2) extended and (3) new features. We start by deriving three base features from key LVC detection measures as described by previous work – GT95, DJ96 and SFN04. As suggested in the previous section, we can make alternate formulations of the past work, such as to discard a pre-filtering step (*i.e.* filtering of constructions that do not include the top three most frequent prepositions). These measures make up the extended feature set. The third set of features are new and have not been used for LVC identification before. These include features that further model the influence of context (*e.g.* prepositions after the object) in LVC detection.

3.1 Base Features

These features are based on the original previous work discussed in Section 2, but have been adapted to give a numeric score. We use the initials of the original authors without year of publication to denote our derived base features.

Recall that the aim of the original GT95 and DJ96 formulae is to rank the possible support verbs given a deverbal noun. As each of these formulae contain a function which returns a numeric score inside the $\arg\max_v$, we use these functions as two of our base features:

$$\text{GT}(v, n) = g(v, n) \quad (4)$$

$$\text{DJ}(v, n) = f(v, n) \sum_n f(v, n) \quad (5)$$

The SFN04 measure can be used without modification as our third base feature, and it will be referred to as SFN for the remainder of this paper.

3.2 Extended Features

Since Grefenstette and Teufel indicated that the filtering step might not be necessary, *i.e.*, $f(v, n)$ may be used instead of $g(v, n)$, we also have the following extended feature:

$$\text{FREQ}(v, n) = f(v, n) \quad (6)$$

In addition, we experiment with the reverse process for the DJ feature, *i.e.*, to replace $f(v, n)$ in the function for DJ with $g(v, n)$, yielding the following extended feature:

$$\text{DJ-FILTER}(v, n) = g(v, n) \sum_n g(v, n) \quad (7)$$

In Grefenstette and Teufel's experiments, they used the top three prepositions for filtering. We further experiment with using all possible prepositions.

3.3 New Features

In our new feature set, we introduce features that we feel better model the v and n components as well as their joint occurrences v-n. We also introduce features that model the v-n pair's context, in terms of deverbal counts, derived from our understanding of LVCs.

Most of these new features we propose are not good measures for LVC detection by themselves. However, the additional evidence that they give can be combined with the base features to create a better composite classification system.

Mutual information: We observe that a verb v and a deverbal noun n are more likely to appear in verb-object pairs if they can form a LVC. To capture this evidence, we employ mutual information to measure the co-occurrences of a verb and a noun in verb-object pairs. Formally, the mutual information between a verb v and a deverbal noun n is defined as

$$I(v,n) = \log_2 \frac{P(v,n)}{P(v)P(n)}, \tag{8}$$

where $P(v,n)$ denotes the probability of v and n constructing verb-object pairs. $P(v)$ is the probability of occurrence of v and $P(n)$ represents the probability of occurrence of n. Let $f(v,n)$ be the frequency of occurrence of the verb-object pair v-n and N be the number of all verb-object pairs in the corpus. We can estimate the above probabilities using their maximum likelihood estimates: $P(v,n) = \frac{f(v,n)}{N}$, $P(v) = \frac{\sum_n f(v,n)}{N}$ and $P(n) = \frac{\sum_v f(v,n)}{N}$.

However, $I(v,n)$ only measures the local information of co-occurrences between v and n. It does not capture the global frequency of verb-object pair v-n, which is demonstrated as effective by Dras and Johnson (1996). As such, we need to combine the local mutual information with the global frequency of the verb-object pair. We thus create the following feature, where the log function is used to smooth frequencies:

$$\text{MI-LOGFREQ} = I(v,n) \times \log_2 f(v,n) \tag{9}$$

Deverbal counts: Suppose a verb-object pair v-n is a LVC and the object n should be a deverbal noun. We denote v' to be the verbalized form of n. We thus expect that v-n should express the same semantic meaning as that of v'. However, verb-object pairs such as "have time" and "have right" in English scored high by the DJ and MI-LOGFREQ measures, even though the verbalized form of their objects, *i.e.*, "time" and "right", do not express the same meaning as the verb-object pairs do. This is corroborated by Grefenstette and Teufel claim that if a verb-object pair v-n is a LVC, then n should share similar properties with v'. Based on our empirical analysis on the corpus using a small subset of LVCs, we believe that:

1. The frequencies of n and v' should not differ very much, and

2. Both frequencies are high given the fact that LVCs occur frequently in the text.

The first observation is true in our corpus where light verb and verbalized forms are freely interchangable in contexts. Then, let us denote the frequencies of n and v' to be $f(n)$ and $f(v')$ respectively. We devise a novel feature based on the hypotheses:

$$\frac{\min(f(n), f(v'))}{\max(f(n), f(v'))} \times \min(f(n), f(v')) \tag{10}$$

where the two terms correspond to the above two hypotheses respectively. A higher score from this metric indicates a higher likelihood of the compound being a LVC.

Light verb classes: Linguistic studies of light verbs have indicated that verbs of specific semantic character are much more likely to participate in LVCs (Wang, 2004; Miyamoto, 2000; Butt, 2003; Bjerre, 1999). Such characteristics have been shown to be cross-language and include verbs that indicate (change of) possession (Danish *give*, to give, direction (Chinese *guan diao* to switch off), aspect and causation, or are thematically incomplete (Japanese *suru*, to do). As such, it makes sense to have a list of verbs that are often used lightly. In our work, we have predefined a light verb list for our English experiment as exactly the following seven verbs: "do", "get", "give", "have", "make", "put" and "take", all of which have been studied as light verbs in the literature. We thus define a feature that considers the verb in the verb-object pair: if the verb is in the predefined light verb list, the feature value is the verb itself; otherwise, the feature value is another default value.

One may ask whether this feature is necessary, given the various features used to measure the frequency of the verb. As all of the other metrics are corpus-based, they rely on the corpus to be a representative sample of the source language. Since we extract the verb-object pairs from the Wall Street Journal section of the Penn Treebank, terms like "buy", "sell", "buy share" and "sell share" occur frequently in the corpus that verb-object pairs such as "buy share" and "sell share" are ranked high by most of the measures. However, "buy" and "sell" are not considered as light verbs. In addition, the various light verbs have different behaviors. Despite their lightness, different light verbs combined with the same noun complement often gives different semantics, and hence affect the lightness of the verb-object pair. For example, one may say that "make copy" is lighter than "put copy". Incorporating this small amount of linguistic knowledge into our corpus-based framework can enhance performance.

Other features: In addition to the above features, we also used the following features: the determiner before the object, the adjective before the object, the identity of any preposition immediately following the object, the length of the noun object (if a phrase) and the number of words between the verb and its object. These features did not improve performance significantly, so we have omitted a detailed description of these features.

4 Evaluation

In this section, we report the details of our experimental settings and results. First, we show how we constructed our labeled LVC corpus, used as the gold standard in both training and testing under cross validation. Second, we describe the evaluation setup and discuss the experimental results obtained based on the labeled data.

4.1 Data Preparation

Some of the features rely on a correct sentence parse. In order to minimize this source of error, we employ the Wall Street Journal section in the Penn Treebank, which has been manually parsed by linguists. We extract verb-object pairs from the Penn Treebank corpus and lemmatize them using WordNet's morphology module. As a filter, we require that a pair's object be a deverbal noun to be considered as a LVC. Specifically, we use WordNet to check whether a noun has a verb as one of its derivationally-related forms. A total of 24,647 candidate verb-object pairs are extracted, of which 15,707 are unique.

As the resulting dataset is too large for complete manual annotation given our resources, we sample the verb-object pairs from the extracted set. As most verb-object pairs are not LVCs, random sampling would provide very few positive LVC instances, and thus would adversely affect the training of the classifier due to sparse data. Our aim in the sampling is to have balanced numbers of potential positive and negative instances. Based on the 24,647 verb-object pairs, we count the corpus frequencies of each verb v and each object n, denoted as $f(v)$ and $f(n)$. We also calculate the DJ score of the verb-object pair $\mathrm{DJ}(v, n)$ by counting the pair frequencies. The data set is divided into 5 bins using $f(v)$ on a linear scale, 5 bins using $f(n)$ on a linear scale and 4 bins using $\mathrm{DJ}(v, n)$ on a logarithmic scale.[1] We cross-multiply these three factors to generate $5 \times 5 \times 4 = 100$ bins. Finally, we uniformly sampled 2,840 verb-object pairs from all the bins to construct the data set for labeling.

4.2 Annotation

As noted by many linguistic studies, the verb in a LVC is often not completely vacuous, as they can serve to emphasize the proposition's aspect, its argument's semantics (*cf.*, θ roles) (Miyamoto, 2000), or other function (Butt and Geuder, 2001). As such, previous computational research had proposed that the "lightness" of a LVC might be best modeled as a continuum as opposed to a binary class (Stevenson et al., 2004). We have thus annotated for two levels of lightness in our annotation of the verb-object pairs. Since the purpose of the work reported here is to flag all such constructions, we have simplified our task to a binary decision, similar to most other previous corpus-based work.

A website was set up for the annotation task, so that annotators can participate interactively. For each selected verb-object pair, a question is constructed by displaying the sentence where the verb-object pair is extracted, as well as the verb-object pair itself. The annotator is then asked whether the presented verb-object pair is a LVC given the context of the sentence, and he or she will choose from the following options: (1) Yes,

[1]Binning is the process of grouping measured data into data classes or histogram bins.

(2) Not sure, (3) No. The following three sentences illustrate the options.

(1) **Yes** – A Compaq Computer Corp. spokeswoman said that the company hasn't *made* a *decision* yet, although "it isn't under active consideration."

(2) **Not Sure** – Besides money, criminals have also used computers to steal secrets and intelligence, the newspaper said, but it *gave* no more *details*.

(3) **No** – But most companies are too afraid to *take* that *chance*.

The three authors, all natural language processing researchers, took part in the annotation task, and we asked all three of them to annotate on the same data. In total, we collected annotations for 741 questions. The average correlation coefficient between the three annotators is $r = 0.654$, which indicates fairly strong agreement between the annotators. We constructed the gold standard data by considering the median of the three annotations for each question. Two gold standard data sets are created:

- **Strict** – In the strict data set, a verb-object pair is considered to be a LVC if the median annotation is 1.
- **Lenient** – In the lenient data set, a verb-object pair is considered to be a LVC if the median annotation is either 1 or 2.

Each of the strict and lenient data sets have 741 verb-object pairs.

4.3 Experiments

We have two aims for the experiments: (1) to compare between the various base features and the extended features, and (2) to evaluate the effectiveness of our new features.

Using the Weka data mining toolkit (Witten and Frank, 2000), we have run a series of experiments with different machine learning algorithms. However, since our focus of the experiments is to determine which features are useful and not to evaluate the machine learners, we report the results achieved by the best single classifier without additional tuning, the random forest classifier (Breiman, 2001). Stratified ten-fold cross-validation is performed. The evaluation criteria used is the F_1-measure on the LVC class, which is defined as

$$F_1 = \frac{2PR}{P+R}, \qquad (11)$$

where P and R are the precision and recall for the LVC class respectively.

4.3.1 Base and Extended Features

Regarding the first aim, we make the following comparisons:

- GT (top 3 prepositions) versus GT (all prepositions) and FREQ
- DJ versus DJ-FILTER (top 3 prepositions and all prepositions)

Feature	Strict	Lenient
GT (3 preps)	0.231	0.163
GT (all preps)	0.272	0.219
FREQ	0.289	0.338
DJ	0.491	0.616
DJ-FILTER (3 preps)	0.433	0.494
DJ-FILTER (all preps)	0.429	0.503
SFN	0.000	0.000

Table 1: F_1-measures of base features and extended features.

We first present the results for the base features and the extended features in Table 1. From these results, we make the following observations:

- Overall, DJ and DJ-FILTER perform better than GT and FREQ. This is consistent with the results by Dras and Johnson (1996).
- The results for both GT/FREQ and DJ show that filtering using preposition does not impact performance significantly. We believe that the main reason for this is that the filtering process causes information to be lost. 163 of the 741 verb-object pairs in the corpus do not have a preposition following the object and hence cannot be properly classified using the features with filtering.
- The SFN metric does not appear to work with our corpus. We suspect that it requires a far larger corpus than our corpus of 24,647 verb-object pairs to work. Stevenson *et al.* (2004)

have used a corpus whose estimated size is at least 15.7 billion, the number of hits returned in a Google search for the query "the" as of February 2006. The large corpus requirement is thus a main weakness of the SFN metric.

4.3.2 New Features

We now evaluate the effectiveness of our class of new features. Here, we do not report results of classification using only the new features, because these features alone are not intended to constitute a stand-alone measure of the lightness. As such, we evaluate these new features by adding them on top of the base features. We first construct a full feature set by utilizing the base features (GT, DJ and SFN) and all the new features. We chose not to add the extended features to the full feature set because these extended features are not independent to the base features. Next, to show the effectiveness of each new feature individually, we remove it from the full feature set and show the performance of classifier without it.

Feature(s)	Strict	Lenient
GT (3 preps)	0.231	0.163
DJ	0.491	0.616
SFN	0.000	0.000
GT (3 preps) + DJ + SFN	0.537	0.676
FULL	**0.576**	**0.689**
- MI-LOGFREQ	0.545	0.660
- DEVERBAL	0.565	0.676
- LV-CLASS	0.532	0.640

Table 2: F_1-measures of the various feature combinations for our evaluation.

Table 2 shows the resulting F_1-measures when using various sets of features in our experiments.[2] We make the following observations:

- The combinations of features outperform the individual features. We observe that using individual base features alone can achieve the highest F_1-measure of 0.491 on the strict data set and 0.616 on the lenient data set respectively. When applying the combination of all base features, the F_1-measures on both data sets increased to 0.537 and 0.676 respectively.

  Previous work has mainly studied individual statistics in identifying LVCs while ignoring the integration of various statistics. The results demonstrate that integrating different statistics (*i.e.* features) boosts the performance of LVC identification. More importantly, we employ an off-the-shelf classifier without special parameter tuning. This shows that generic machine learning methods can be applied to the problem of LVC detection. It provides a sound way to integrate various features to improve the overall performance.

- Our new features boost the overall performance. Applying the newly proposed features on top of the base feature set, *i.e.*, using the full feature set, gives F_1-measures of 0.576 and 0.689 respectively (shown in bold) in our experiments. These yield a significant increase ($p < 0.1$) over using the base features only. Further, when we remove each of the new features individually from the full feature set, we see a corresponding drop in the F_1-measures, of 0.011 (deverbal counts) to 0.044 (light verb classes) for the strict data set, and 0.013 (deverbal counts) to 0.049 (light verb classes) for the lenient data set. It shows that these new features boost the overall performance of the classifier. We think that these new features are more task-specific and examine intrinsic features of LVCs. As such, integrated with the statistical base features, these features can be used to identify LVCs more accurately. It is worth noting that light verb class is a simple but important feature, providing the highest F_1-measure improvement compared to other new features. This is in accordance with the observation that different light verbs have different properties (Stevenson et al., 2004).

5 Conclusions

Multiword expressions (MWEs) are a major obstacle that hinder precise natural language processing (Sag et al., 2002). As part of MWEs, LVCs remain least explored in the literature of computational linguistics. Past work addressed the problem of automatically detecting LVCs by employing single statistical measures. In this paper, we experiment

[2] For the strict data set, the base feature set has a precision and recall of 0.674 and 0.446 respectively, while the full feature set has a precision and recall of 0.642 and 0.523 respectively. For the lenient data set, the base feature set has a precision and recall of 0.778 and 0.598 respectively, while the full feature set has a precision and recall of 0.768 and 0.624 respectively.

with identifying LVCs using a machine learning framework that integrates the use of various statistics. Moreover, we have extended the existing statistical measures and established new features to detect LVCs.

Our experimental results show that the integrated use of different features in a machine learning framework performs much better than using any of the features individually. In addition, we experimentally show that our newly-proposed features greatly boost the performance of classifiers that use base statistical features. Thus, our system achieves state-of-the-art performance over previous approaches for identifying LVCs. As such, we suggest that future work on automatic detection of LVCs employ a machine learning framework that combines complementary features, and examine intrinsic features that characterize the local context of LVCs to achieve better performance.

While we have experimentally showed the effectiveness of the proposed framework incorporating existing and new features for LVC detection on an English corpus, we believe that the features we have introduced are generic and apply to LVC detection in other languages. The reason is three-fold:

1. Mutual information is a generic metric for measuring co-occurrences of light verbs and their complements. Such co-occurrences are often an obvious indicator for determining light verbs because light verbs are often coupled with a limited set of complements. For instance, in Chinese, directional verbs, such as *xia* (descend) and *dao* (reach), which are often used lightly, are often co-located with a certain class of verbs that are related to people's behaviors.

2. For LVCs with noun complements, most of the semantic meaning of a LVC is expressed by the object. This also holds for other languages, such as Chinese. For example, in Chinese, *zuo xuanze* (make a choice) and *zuo jueding* (make a decision) has the word *zuo* (make) acting as a light verb and *xuanze* (choice) or *jueding* (decision) acting as a deverbal noun (Wang, 2004). Therefore, the feature of deverbal count should also be applicable for other languages.

3. It has been observed that in many languages, light verbs tend to be a set of closed class verbs. This allows us to use a list of predefined verbs that are often used lightly as a feature which helps distinguish between light and non-light verbs when used with the same noun complement. The identity of such verbs has been shown to be largely independent of language, and corresponds to verbs that transmit information about possession, direction, aspect and causation.

References

T. Bjerre. 1999. Event structure and support verb constructions. In *4th Student Session of European Summer School on Logic, Language and Information 1999*. Universiteit Utrecht Press, Aug, 1999.

L. Breiman. 2001. Random forests. *Machine Learning*, 45(1):5–32, Oct, 2001.

M. Butt and W. Geuder. 2001. On the (semi)lexical status of light verbs. In *Semi-lexical Categories*, pages 323–370. Mouton de Gruyter.

M. Butt. 2003. The light verb jungle. In *Workshop on Multi-Verb Constructions*.

M. Dras and M. Johnson. 1996. Death and lightness: Using a demographic model to find support verbs. In *5th International Conference on the Cognitive Science of Natural Language Processing*.

A. Fazly, R. North, and S. Stevenson. 2005. Automatically distinguishing literal and figurative usages of highly polysemous verbs. In *ACL 2005 Workshop on Deep Lexical Acquisition*, pages 38–47.

G. Grefenstette and S. Teufel. 1995. A corpus-based method for automatic identification of support verbs for nominalizations. In *EACL '95*.

T. Miyamoto. 2000. *The Light Verb Construction in Japanese. The role of the verbal noun.* John Benjamins.

I. Sag, T. Baldwin, F. Bond, A. Copestake, and D. Flickinger. 2002. Multiword expressions: A pain in the neck for NLP. In *Lecture Notes in Computer Science*, volume 2276, Jan, 2002.

S. Stevenson, A. Fazly, and R. North. 2004. Statistical measures of the semi-productivity of light verb constructions. In *2nd ACL Workshop on Multiword Expressions: Integrating Processing*, pages 1–8.

L. Wang. 2004. A corpus-based study of mandarin verbs of doing. *Concentric: Studies in Linguistics*, 30(1):65–85, Jun, 2004.

I. Witten and E. Frank. 2000. *Data Mining: Practical machine learning tools with Java implementations*. Morgan Kaufmann.