International Journal of Knowledge and Language Processing
Volume 9, Number 2, **2018**

www.ijklp.org
KLP International ©2018 ISSN 2191-2734
pp.33—46

Supervised Word Sense Disambiguation with Frame-based Constructional Features: a pilot study of *fán* 煩 "to annoy/be annoying/be annoyed"*

Mingyu Wan[1] and Meichun Liu[2]
[1,2]Department of Linguistics and Translation
City University of Hong Kong
83, Tat Chee Avenue Kowloon, Hong Kong SAR
{mywan4, meichliu}@cityu.edu.hk

Received July 2018; revised October 2018

ABSTRACT. *Studies of Word Sense Disambiguation have made extensive use of unstructured contextual features for disambiguating polysemous words, while this paper aims to testify the validity of adopting linguistically-motivated and semantically-encoded features for classifying word senses. It conducts an innovative pilot study on supervised Word Sense Disambiguation of a Chinese polysemous verb fán 煩 "to annoy/be annoying/be annoyed" via the use of frame-based constructional (FC) features. Experimental results have shown significant improvement of using FC over the baseline sets, with the weight average* $F_{\Delta max}$ *attains to 0.312. Besides, the noun phrase feature set also shows impressive performance compared to uni-grams and bi-grams, which tends to imply a close relation between verb meaning and core arguments. The promising results have proved the great discriminativeness of FC for sense disambiguation and suggest a possible alternation of employing/combining deep linguistic resources for Natural Language Processing applications in future.*

Keywords: Word Sense Disambiguation, Frame-based Constructional Features, Supervised Machine Learning, Polysemous Emotion Verb *fán*, Lexicalization Pattern

* An earlier version of the study was presented at the 2018 CLSW conference in Tai Wan and is developed into a fuller paper here, with the original dataset expanded from 200 sentences to 500 sentences. The data used in the study is from Mandarin VerbNet, an online semantic lexicon of Mandarin verbs.

1. **Introduction.** Lexical semantic ambiguity, also known as Polysemy[1] in Linguistics, has been a problematic issue of Natural Language Processing (NLP) applications, for it may cause inferior performance to tasks that rely largely on the accurate identification of word meaning, such as in Machine Translation, Information Retrieval, automatic part-of-speech tagging and syntactic parsing [1]. Word Sense Disambiguation (WSD) has therefore arisen to tackle the problem by referring to the context of the polysemous word. In the following paragraphs, we will firstly give a general review on the related studies and then highlight the uniqueness and significance of our work.

Studies of WSD can be traced back earliest to 1950s [2] when Kaplan adopted a few simple rules to infer the word sense. Since then the rule-based method had been dominant in this subject for a few decades. Recently, with the prevalence of digitalized techniques, two more methods are commonly adopted in the current studies: dictionary-based (see *e.g.* [3, 4]) and statistics-based. Since 1990s, Machine Learning has become popular in the state-of-the-arts, and it mainly includes three methods: supervised Machine Learning (see *e.g.* [5, 6]), unsupervised Machine Learning (see *e.g.* [7]), and semi-supervised Machine Learning (see *e.g.* [8]). The supervised method is to predict the verb class with external labels, that is, to train the Machine Learning models on representative features of the target class on the basis of a pre-categorized dataset. The supervised mode is claimed to achieve the best performance out of a relatively small set of structured data, which however requires manual annotation in the data pre-processing; in contrast, unsupervised Machine Learning is a method without any external pre-labels, and it is suitable for clustering tasks on big data. However, it is computationally more demanding in terms of feature selection and clustering algorithms; semi-supervised Machine Learning is a hybrid approach between the two. In this paper, we adopt the supervised Machine Learning for WSD with the use of a small set of semantically-annotated data (500 sentences[2] containing the polysemous verb *fán* 煩) from the well-structured semantic online lexicon–Mandarin VerbNet (MV) [9]. This database has enabled us to attest the validity of using linguistically-enriched context for WSD. Details of the approach to the linguistic analysis of the verb and the semantic annotation will be introduced in section 2.

In terms of the choice of representative features in the context, past studies (see *e.g.* [10, 11]) have extensively adopted pure lexical features, such as n-grams of words and characters (see *e.g.* [12]), and word-to-vectors [13]. A few researches are based on similarity measurement (see *e.g.* [14]). Few attempts are found on using deep linguistic features. Besides, many relevant studies are focused on English polysemous words, WSD of Chinese polysemous words has far less been dealt with. One reason is the semantic heterogeneity of Chinese words (esp. monosyllabic words), and another is the difficulty of getting access to a semantic resource of Chinese which is systematic and fine-grained

[1] Polysemy is the association of one word with two or more distinct but related meanings, such as the case of '*bank*'. In WordNet, it has 10 meanings as a noun and 8 meanings as a verb.

[2] The dataset has been expanded to 500 instances compared to the original 200 instances in the preliminary experiment. The new results generally conform to the original findings, but with minor differences (see detailed results in section 4).

enough. As a pilot study, this paper takes an innovative trial by employing frame-based constructional features (see *e.g.* [15-18]) for supervised WSD on the Chinese polysemous verb *fán* 煩 "to annoy/be annoying/be annoyed", with the aims to 1) introduce the frame-based constructional approach to the resolution of lexical semantic ambiguity; 2) testify the discriminative power of frame-based semantic-to-syntactic encodings for WSD, in comparison to other basic features, such as n-grams of neighboring words and phrases; 3) highlight the significance of using deep linguistic interpretations for improving NLP tasks in the long run.

To fulfill the above research purposes, this paper will give a detailed account of the frame-based constructional (FC) approach to emotion verbs by focusing on the Chinese verb *fán* 煩 in section 2 and conduct eight supervised WSD experiments on four sets of features with two classifiers in section 3. In section 4, the results will be displayed and discussed with data interpretation and analysis. Conclusions will be given in section 5.

2. **The FC Approach to Emotion Verbs in Chinese.**

2.1. **What is Frame-based Constructional Approach?** In this paper, the Frame-based Constructional (FC) features we use for the WSD experiment are based the frame-based constructional annotations in the MV verbal lexicon. The FC approach to Mandarin verbs is proposed by Liu [15–18], which is a hybrid approach for annotating Mandarin verbal lexicon, by incorporating tenets of Frame semantics [19, 20], Construction Grammar [21] and Cognitive Grammar [22].

According to Frame Semantics, the meaning of a verb can be defined only in relation to a structured background of eventive knowledge and experiences. The background frame is shared by semantically related lemmas that can be best described and unified within a set of frame-specific participant roles, *i.e.* frame elements. Take two basic domains for example, Financial Transaction verbs share a background frame consisting of core elements such as Buyer, Seller, Goods, Money, as in "[Lee]$_{\text{Buyer}}$ BOUGHT [a book]$_{\text{Goods}}$ from [Aby]$_{\text{Seller}}$ with [2 dollars]$_{\text{Money}}$". Emotion verbs are usually set in a frame of Experiencer, Stimulus and other related elements, as in "[Money]$_{\text{Stimulus}}$ PLEASES [John]$_{\text{Experiencer}}$". In addition to the frame-based conceptualization, a commonly held belief in lexical semantic studies is that the meaning of a verb is manifested in syntactic realizations [23]. Under this premise, verbal meanings can only be distinguished if they are syntactically detectable. Expanding upon the frame-verb relation by integrating the syntactic-to-semantic notion that verbal meanings can be distinguished by the help of their formal behaviors, the FC approach refines the semantic notion of frames with the aid of syntactic constraints from Construction Grammar. A construction is defined as a basic form-meaning mapping template that can be instantiated with a semantically compatible verb as an instance of construction realization. Constructions and verbs, both as meaning-bearing units, go hand-in-hand in defining the semantics of argument realizations characteristic of a given background frame.

The following example of the placement verb 放 *fang* 'to put/place' is shown below to illustrate the FC approach, together with the semantic annotation:

Example 2.1. 他 把 書 放 在 房間裏

tā *bǎ* *shū* *fàng* *zài* *fángjiān-lǐ*

3p.sg BA book put at room-inside

'She put the book in the room.'

The semantic annotation of the above example with the FC tags is shown below:

→ [他]Placer [把]*BA [書]Figure [放]PLACEMENT [在]*Locative_mkr [房間裏]Ground

As shown above, the different semantic tags of defining the placement verb *fáng* 放 are categorized into three kinds of information: the frame (all capitalized), the core frame elements (initial-letter capitalized), as well as the constructional markers (marked by asterisks) of the verb, as summarized below:

- **Frame:** PLACEMENT
- **Core frame elements:** Placer, Figure, Ground_loc
- **Construction markers:** *BA, *Locative_mkr

Frame elements are the verb-specific core and non-core elements (participant roles) that can profile the verb meaning on the basis of Fillmore's schematic theory that verb senses are anchored in frames. Construction markers are the salient syntactic indicators that are closely associated to the verbs, and they are based on Levin's alternation-based approach and Goldberg's Construction Grammar indicating the close relation between verb classes and syntactic constructions. The following section will focus on the introduction of Liu's analysis of Emotion verbs [16] to specify the FC approach to the domain of Emotion verbs.

2.2. **FC to Emotion Verbs and the Special Lexicalization Patterns.** In the study of lexical semantics, Emotion (psycho-logical or psych) verb is one of the most fundamental and appealing topics to linguists. For example, Talmy [24] introduced the dichotomy of emotional valence in terms of subject selection (Stimulus as subject vs. Experiencer as subject). Valin [25] proposed one crucial notion of "Effector" in addition to Talmy's basic theory, which underpins the property of a volitional and acting instigator involved in an inchoative event (causing change to the affected). On their basis, Liu [16] comprehensively studied the five-way distinctions of lexicalization pattern of emotion verbs, as shown in the following examples.

Example 2.2. *stimulus-as-subject + transitive (Liu [16: example 54d])*

這個 問題 煩 了 我 三天三夜

zhè-gè *wèntí* *fán* *le* *wǒ* *sāntiānsānyè*

this-CL problem annoy ASP 1p.sg three-days-and-nights

'This problem has bothered me for three days and nights.'

The semantic annotation of the above example with the FC tags is shown below:

→ [這個問題]Stimulus [煩]STIM-SUBJ了[我]Experiencer[三天三夜]Duration

Example 2.3. *stimulus-as-subject + intransitive (Liu [16: example 54c])*

明天　　的　　考試　　很　　煩

míngtiān　de　kǎoshì　hěn　fán

tomorrow　DE　exam　DEG　annoy

'The exam tomorrow is very annoying.'

The semantic annotation of the above example with the FC tags is shown below:

→ [明天的考試]$_{\text{Stimulus}}$ [很]$_{\text{*Degree}}$[煩]$_{\text{STIM-SUBJ}}$

Example 2.4. *experiencer-as-subject + transitive (Liu [16: example 54b])*

你　　在　　煩　　什麼？

nǐ　zài　fán　shíme

2p.sg PROG　bother　what

'What are you bothered of?'

The semantic annotation of the above example with the FC tags is shown below:

→ [你]$_{\text{Experiencer}}$在[煩]$_{\text{EXP-SUBJ}}$[什麼]$_{\text{Target}}$?

Example 2.5. *experiencer-as-subject + intransitive (Liu [16: example 54a])*

我　　好　　煩　　啊!

wǒ　hǎo　fán　ā!

1p.sg　DEG　bother　EXC

'I am so much bothered!'

The semantic annotation of the above example with the FC tags is shown below:

→ [我]$_{\text{Experiencer}}$[好]$_{\text{*Degree}}$[煩]$_{\text{EXP-SUBJ}}$啊!

Example 2.6. *affector-as-subject (Liu [16: example 53c])*

老闆　　一直　　煩　　他

lǎopǎn　yīzhí　fán　tā

boss　always　trouble　3p.sg

'The boss has been bothering him'

The semantic annotation of the above example with the FC tags is shown below:

→ [老闆]$_{\text{Affector}}$一直[煩]$_{\text{BOTHER}}$[他]$_{\text{Affectee}}$!

2.3. **The Tripartite Senses of *fán* 煩.** On the basis of the five-way distinctions of emotion verbs in Liu [16], we find from corpus data that the verb *fán* is polysemous in that it can denote three salient possible meanings: to annoy (S1), be annoying (S2) or be annoyed (S3). In MV, the three senses are tagged as BOTHER (affector-as-subject, with volition of the agentive emotion), STIM-SUBJ (stimulus-as-subject, with the causative stimulus for provoking an emotion) and EXP-SUBJ (experiencer-as-subject, with the direct emotion of the experiencer) respectively in the dataset. The following examples of *fán* 煩 are used to show some of the polysemous cases:

Example 2.7. *Two possible interpretations, with semantic annotations:*

他　　真的　　好　　煩!

tā zhēnde hǎo fán

3p.sg really DEG annoy

S2: 'He is really annoying!'→[他]$_{\text{Stimulus}}$真的[好]$_{\star\text{Degree}}$ [煩]$_{\text{STIM-SUBJ}}$

S3: 'He felt so annoyed!' →[他]$_{\text{Experiencer}}$真的[好]$_{\star\text{Degree}}$ [煩]$_{\text{EXP-SUBJ}}$

Example 2.8. *Three possible interpretations, with semantic annotations:*

數學 老師 煩 死 他 了!

shùxué lǎoshī fán sǐ tā le

Math teacher annoy DEG 3p.sg LE

S1: 'The Math teacher has annoyed him a lot (with volition)!'

→[數學老師]$_{\text{Affector}}$ [煩]$_{\text{BOTHER}}$[死]$_{\star\text{Degree}}$[他]$_{\text{Affectee}}$了

S2: 'The Math teacher is so annoying for him (without volition)!'

→[數學老師]$_{\text{Stimulus}}$ [煩]$_{\text{STIM-SUBJ}}$[死]$_{\star\text{Degree}}$[他]$_{\text{Experiencer}}$了

S3: 'The Math teacher is so annoyed of him!'

→[數學老師]$_{\text{Experiencer}}$ [煩]$_{\text{EXP-SUBJ}}$[死]$_{\star\text{Degree}}$[他]$_{\text{Target}}$了

Example 2.9. *Two possible interpretations, with semantic annotations:*

他 煩 你 了?

tā fán nǐ le

3p.sg annoy 2.sg LE

S1: 'Has he bothered you?' →[他]$_{\text{Affector}}$ [煩]$_{\text{BOTHER}}$[你]$_{\text{Affectee}}$了?

S3: 'Is he annoyed of you?' →[他]$_{\text{Experiencer}}$ [煩]$_{\text{EXP-SUBJ}}$[你]$_{\text{Target}}$了?

The examples (2.7-2.9) have proved the polysemous property of the verb *fán* as it can be ambiguous with two or three possible meanings (S1-3). For each interpretation of the ambiguous verb, it shows distinct construction patterns correlated to that particular sense, as the multiple meanings are anchored in their core Frame Elements and constructions. For example: S1 (to annoy) has a distinct construction pattern (CP) of [Affector]-[V]-[Affectee], where [V] represents the prediction verb; S2 (be annoying) has a distinct CP of [Stimulus]-[*Degree]-[V]; and S3 (be annoyed) has a distinct CP of [Experiencer]-[*Degree]-[V]. On the basis of Liu's analysis of emotion verbs [16], as well as the form-meaning mapping principle [19, 21, 23], this study is enabled to borrow a gold standard training data which is encoded with rich semantic information for the supervised WSD experiments, as will be shown in Section 3 and 4.

3. **Experiment Setup.**

3.1. **Data and Tools.** The training and testing data is from the database of MV, in which there are 500 sentences that contain *fán* 煩 as the main predicate and are well annotated with frame-based constructional information. The annotation process is executed semi-automatically by using the open source editor Atom with specially designed embedding package 'VerbNet Tool' for data management. Manual Validation and Inter

Annotator Agreement are also conducted to ensure the accurateness of the annotation.

In the supervised WSD task, the dataset is experimented in the ten-fold cross validation mode (see [26]) which is an effective way for Machine Learning that deals with a small size of data and to avoid a biased result. The raw sentences before the semantic annotation were retrieved and randomly sampled from the Chinese Gigaword (LDC2003T09) with a normal distribution. Lexical n-gram collocation features of the polysemous verb are automatically extracted from the raw sentences by self-programs run on PyCharm Community [27]. Syntactic collocation features of the main verb are also automatically extracted from parsed trees by Stanford Parser [28] with manual adjustment. The Machine Learning tasks are carried out by using the open source data mining software WEKA [29], which provides a collection of several state-of-the-art Machine Learning algorithms.

3.2. **Feature sets.** Four feature sets have been tested in the WSD experiments and they are:

Uni-gram: +1 and －1 window size in character unit of the ambiguous word. Uni-gram represents the one-character context of the ambiguous verb. For example, in the sentence "你少烦我", the Uni-gram characters would be "你", "少" and "我". For the 500 sentences of *fán*, there are 1006 types of Uni-grams in total, which constitute the Uni-gram feature set with 1007 attributes (adding one attribute of verb sense) and 500 instances (samples) in the WSD experiment, serving as one baseline set for comparison.

Bi-gram: +2 and －2 window size in character unit of the ambiguous word. Bi-gram represents the two-character context of the ambiguous verb. For example, in the sentence "你少烦我", the Bi-grams would be "你少"and "我 Ø", where Ø represents a null existence for forming a bi-gram. For the 500 sentences of *fán*, there are 3249 types of Bi-grams in total, which constitute the Bi-gram feature set with 3250 attributes and 500 instances in the WSD experiment, serving as another baseline set for comparison.

Uni-NP: +1 and －1 window size in NP unit of the ambiguous word. Uni-NP represents the one-NP context of the ambiguous verb. For example, in the sentence "你少烦我", the Uni-NP would be "你" and "我". If there is no noun phrase surrounding the main verb, such as in "烦死了", the Uni-NP would be Ø. For the 500 sentences of *fán*, there are 266 types of Uni-NPs in total, which constitute the Uni-NP feature set with 267 attributes and 500 instances in the WSD experiment, serving as the third baseline set for comparison.

FC: the frame-based constructional context of the ambiguous word. FC represents the semantic-syntactic context of the ambiguous verb, as shown in the examples of section 2. For example, in the annotated sentence "[你]$_{\text{Affector}}$少[烦]$_{\text{BOTHER}}$[我]$_{\text{Affectee}}$", the FC would be "[Affector]" and "[Affectee]". For the 500 annotated sentences of *fán*, there are 21 types of FC features in total, which constitute the FC feature set with 22 attributes and 500 instances in the WSD experiment, serving as the target feature set.

The four kinds of feature sets are converted to attribute tables for the WSD Machine Learning tasks by self-programs, in which the attribute property is Boolean value, *i.e.,* 0 for absence and 1 for presence of one particular attribute (feature) in one instance (sentence). These attribute values, together with the pre-categorized labels (sense of the verb), are then fed into the learning models for training and testing with ten-fold cross validation. In the

testing process, the performances of the WSD tasks using the four feature sets are calculated automatically and output by Weka.

3.2. **Classifiers.** The Naïve Bayes (NB) and Sequential Minimal Optimization (SMO) models are adopted as the classifiers for the experiment in this paper. Preliminary experiments on using several state-of-the-art classifiers (e.g. NB, SMO, J48-decision tree, K-Nearest-Neighboring and so on) consistently showed the outstanding performance of NB and SMO compared to the other classifiers. They have been claimed to perform outstandingly well with either data sparseness or the overfitting problem. (see e.g. [26, 30, 31]).

NB: The Naïve Bayes algorithm is most widely adopted in Machine Learning because of its simplicity and fast speed of building the model, yet with impressive performances; it is a probabilistic classifier based on the assumption that all the predicators are mutual independent. In real cases, such assumption is indeed Naïve, but it is still claimed to perform surprisingly well despite of the correlation of the predictors; Li and Jain [30] indicated that NB is good at dealing with the over-fitting problem and the performance of NB improves as the number of features increases. Besides, they found that NB requires only a small number of training data to achieve good performance.

SMO: The Sequential Minimal Optimization algorithm is an advanced SVM (Support Vector Machine) that is realized by John Platt's pairwise classification model [32] which effectively solved the Quadratic Programming (QP) problem by decomposing it into small sequences of minimal optimizations; SMO is also enabled for multi-class classification by using the "one vs. one" algorithm. As proven by many researchers, SVM is a very powerful classifying model that shows significantly better performance than many state-of-the-art classifiers in most cases. Besides, Joachims [33] explained that SVM uses over-fitting protection to ensure its well performance for dealing with features of high dimensionality, and it does not require parameter tuning to achieve high accuracy, as he put it "With their ability to generalize well in high dimensional feature spaces, SVMs eliminate the need for feature selection, making the application of text categorization considerably easier."

3.3. **Evaluation Metric.** In this paper, F-Measure, also known as F-score, is adopted as the metric for evaluating the WSD performance. It is a most widely used formula for calculating classification performance. The following formulae illustrate how F-Measure is defined:

$$\text{Precision: } P=TP/(TP+FP)$$

$$\text{Recall: } R=TP/(TP+FN)$$

$$\text{F-Measure: } F=2PR/(P+R)$$

Where TP = true positives, FP = false positives, FN = false negatives.
More details about the measurement can be found in Manning and Schütze [34].

4. **Results and Discussions.** In the following tables and graphs, S1 stands for sense 1 ('to annoy') which corresponds to the BOTHER frame, S2 stands for sense 2 ('be annoying') which corresponds to the STIM-SUBJ frame, and S3 stands for sense 3 ('be annoyed')

which corresponds to the EXP-SUBJ frame; '*W Avg.*' stands for the weighted average Precision (P), Recall (R) and F-Measure (F); 'F_{Δ}' stands for the F-Measure discrepancy between a pair of feature sets in comparison. '$F_{\Delta max}$' stands for the maximum 'F_{Δ}' among a group of feature sets. The experimental results and discussions are shown in the following subsections:

4.1. **Overall Performance.** This section gives a general comparison among all the four feature sets with the two classifiers for both individual classes (senses) and averaged performance. See the summarized results in terms of F in the following table.

TABLE 1. OVERALL PERFORMANCE OF THE FOUR FEATURE SETS

	Uni-gram		***Bi-gram***		***Uni-NP***		***FC***		$F_{\Delta max}$
	NB	*SMO*	*NB*	*SMO*	*NB*	*SMO*	*NB*	*SMO*	
S1	*0.378*	*0.358*	*0.624*	*0.634*	*0.596*	*0.629*	*0.990*	*1.000**	*0.642*
S2	*0.467*	*0.544*	*0.491*	*0.590*	*0.627*	*0.655*	*0.896*	*0.899**	*0.432*
S3	*0.745*	*0.772*	*0.744*	*0.780*	*0.822*	*0.829*	*0.939**	*0.934*	*0.195*
W Avg.	*0.618*	*0.657*	*0.650*	*0.704*	*0.682*	*0.704*	*0.930**	*0.929*	***0.312***

As shown in Table 1, the FC feature set consistently achieves the best F (marked by *) for each sense class with either NB or SMO, while the worst performance (marked by underlines) falls in the range between Uni-gram and Bi-gram. In terms of the *W Avg.* F, the FC set outperforms the other three feature sets with an outstanding score of 0.930 *via* the use of NB; by contrast, Uni-gram shows the lowest *W Avg*. F of 0.618 *via* the use of NB. The $F_{\Delta max}$ stands for a maximal F discrepancy among a group of figures in comparison, and the *W Avg.* $F_{\Delta max}$ (0.312) shows the averaged F discrepancy out of the three individual senses. In order for a more visualized presentation of the classification results, we use the following Figure for further illustration.

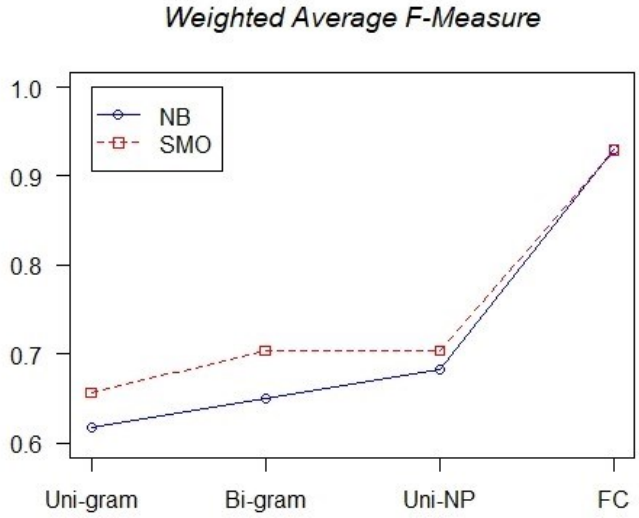

FIGURE 1. THE OVERALL PERFORMANCE OF THE FOUR FEATURE SETS WITH TWO CLASSIFIERS

The result in Figure 1 generally shows a significant improvement of using FC features for WSD over the three baselines, indicating the greater discriminativeness of the

frame-based constructional features than the lexical features for sense classifying, which echoes the "one frame, one sense" principle proposed by Liu [15-18]. It also indicates the usefulness of adopting argument structures (Uni-NP) for WSD, compared to the lexical features, which seems to suggest a positive correlation between argument realization and verb sense. Within the two n-gram feature sets, Bi-gram shows a slight better performance than Uni-gram, which can be explained by the skewed feature space between the two sets. As shown in section 3.1, the feature dimension of Bi-gram is much higher than of Uni-gram (feature dimension: 3250 *vs.* 1007), and both classifiers are good at dealing with high dimension features without the over-fitting problem. The higher feature space of Bigram renders a higher chance of including more indicative features for inferring the correct senses. One additional observation is that the classification performance of each feature set varies to certain degree at different individual verb senses (S1-3). The following subsections will give more specific comparisons from more perspectives.

4.2. **Cross-sense Comparison.** This section gives a comparison of mean F (an averaged F of the four feature sets for each sense of S1-3) across the three senses with the two classifiers. The results are shown in Figure 2 below.

It clearly shows the largely scattered performances of the three different senses for both classifiers: S3 (be annoyed) can be disambiguated most successfully than S1 (to annoy) and S2 (be annoying); S1 is slightly easier to discern compared to S2. In order to find out the possible reasons for causing such differences, Figure 3 is given to show the instance distribution of the three senses in the dataset to account for the performance variation.

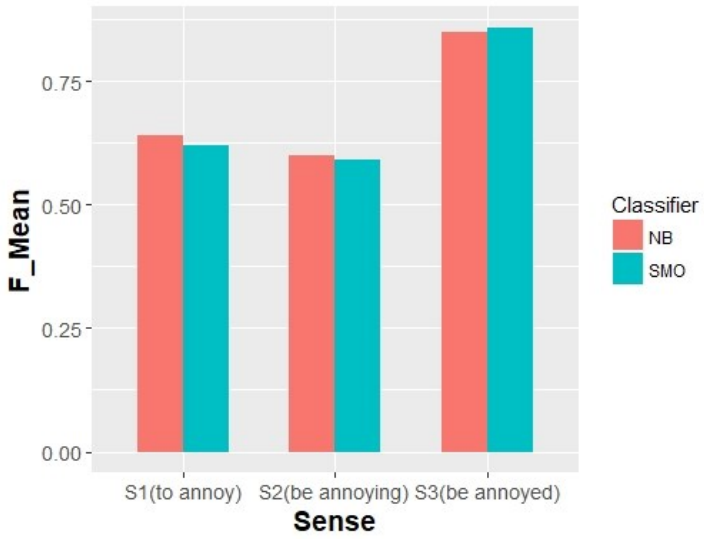

FIGURE 2. THE MEAN F ACROSS THE THREE SENSES

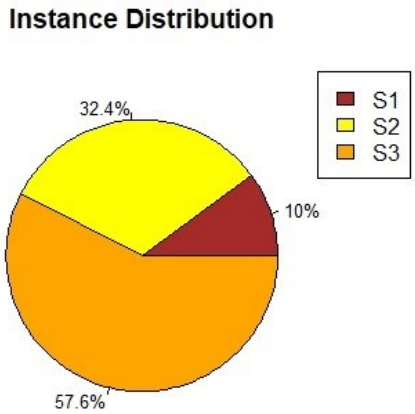

FIGURE 3. INSTANCE DISTRIBUTION OF THE THREE SENSES IN THE DATASET

As clearly shown in Figure 3, S3 accounts for the largest proportion (57.6%) of the 500 instances which ensures a biggest number of training data among the three. With a larger size of training data, the classifiers are more likely to disambiguate the "be annoyed" sense. But interestingly, even though S2 (be annoying) takes up a higher proportion of instances than S1 (to annoy) (32.4% *vs.* 10%), its WSD performance is unexpectedly worse. This can be explained by the more confusing context of the "be annoying" sense with the other senses, as shown in the examples (2.7-2.9). The classifiers are more likely to be confused when labelling this sense. The following subsection will give a fuller analysis for explaining such confusion by focusing on the FC set.

4.3. **Detailed Performance of FC.** This section shows more detailed results of FC in the WSD tasks. The main classification output in terms of P, R, and F is shown in Table 2 and the confusion matrices of using the two classifiers are displayed in Table 3 to help identify the most confusing senses.

TABLE 2. PERFORMANCE OF FC IN TERMS OF P, R AND F

	NB			***SMO***		
	P	*R*	*F*	*P*	*R*	*F*
S1	*0.980*	*1.000*	*0.990*	*1.000*	*1.000*	*1.000*
S2	*0.886*	*0.907*	*0.896*	*0.825*	*0.988*	*0.899*
S3	*0.947*	*0.931*	*0.939*	*0.992*	*0.882*	*0.934*
W Avg.	*0.930*	*0.930*	*0.930*	*0.939*	*0.928*	*0.929*

Results in Table 2 show that 1) S1 tends to perform well in terms of both Precision and Recall for both classifiers, while S3 tends to perform well in terms of Precision and S2 tends to perform well in terms of Recall; 2) S2 performs worst in terms of Precision among the three senses for both classifiers and this causes its low F in average; 3) the two classifiers show equal performance in average with minor discrepancy (F_{Δ}=0.001). The following two confusion matrices in Table 3 is shown below to help to find out the most confusing cases.

TABLE 3. CONFUSION MATRICES OF FC WITH THE TWO CLASSIFIERS

NB				***SMO***			
a	*b*	*c*	← *classified as*	*a*	*b*	*c*	← *classified as*
50	*0*	*0*	*a = S1*	*50*	*0*	*0*	*a = S1*
0	*147*	*15*	*b = S2*	*0*	*160*	*2*	*b = S2*
1	*19*	*268*	*c = S3*	*0*	*34*	*254*	*c = S3*

As shown in Table 3, for the NB classifier, 15 instances of S2 are misclassified as S3, and 19 instances of S3 are misclassified as S2; similarly, for the SMO classifier, 2 instances of S2 are misclassified as S3, and 34 instances of S3 are misclassified as S2. This proves that S2 (be annoying) and S3 (be annoyed) are the most confusing senses among the three, which conforms to the observation in the examples (2.7-2.9). Besides, one instance of S3 is misclassified as S1 for the NB classifier, but this does not happen with the SMO classifier. This may imply a more robust performance of SMO, compared to NB, or a better fitness of SMO with low dimension feature set (22 attributes for FC).

5. **Conclusions.** This paper has innovatively made use of FC features to disambiguate the tripartite senses of the Chinese polysemous verb *fán* 煩 "to annoy/be annoying/be annoyed". Based on the 500 semantically annotated sentences from Mandarin VerbNet with FC features, the experiments have obtained fairly impressive results. It is suggested that carefully encoded semantic information is significantly more effective for discerning word senses than pure lexical context. The findings have also echoed the form-meaning mapping principle of Liu [15–18] that verb meaning is closely associated with argument structures, and it is realized with salient constructional patterns. This indicates a dynamic interaction between lexis, construction and verb sense. To conclude, this paper serves as a good example of exploring deep linguistic encodings for NLP applications and has paved the way for more possibilities in NLP advancement. However, the challenge goes to the difficulty in labeling such FC tags in a fully automatic way, which elicits the need of automating FC labeling in the near future. Besides, one limitation of this study is that the size of the dataset is rather small, which results in the salient performance discrepancy between the baseline sets and the FC set. In the case of larger data, the lexical features are expected to boost performance to certain degree, but the case of FC features still needs further examination. It is still safe to conclude that frame-based constructional features are more discriminative than pure lexical features, at least for predicting verb senses, and the advantage is even obvious when the dataset is small. As implied from the results, the linguistically-enriched tags may become extremely useful when we encounter the problem of data sparseness.

Acknowledgment. This work is partially supported by an internal research grant from City University of Hong Kong (Project No.: 7004742). The authors owe great gratitude to the helpful comments and suggestions of the reviewers of CLSW 2018 on the first version of the paper. We are also thankful to the two students who gave valuable advice to the improvement of our experiment after the presentation of the earlier paper at the conference.

REFERENCES

[1] Navigli, R.: Word sense disambiguation: A survey. *ACM Computing Surveys* 41(2), 10 (2009).

[2] Kaplan, A.: An experimental study of ambiguity and context. *Mechanical Translation* 2(2), 39–46 (1950).

[3] Banerjee, S., Pedersen, T.: An adapted Lesk algorithm for word sense disambiguation using WordNet. In*: International Conference on Intelligent Text Processing and Computational Linguistics*, pp. 136–145. Springer, Berlin, Heidelberg (2002).

[4] Brzeski, V., Reiner, K.: Word sense disambiguation. U.S. Patent Application No. 11/270, 917, (2005).

[5] Ide, N., Jean, V.: Introduction to the special issue on word sense disambiguation: the state of the art. *Computational linguistics* 24(1), 2–40 (1998).

[6] Lee, Y.K., Hwee, T.N., Tee, K.C.: Supervised word sense disambiguation with support vector machines and multiple knowledge sources. In: *Senseval-3: third international workshop on the evaluation of systems for the semantic analysis of text*. (2004).

[7] Yarowsky, D.: Unsupervised word sense disambiguation rivaling supervised methods. In: *Proceedings of the 33rd annual meeting on Association for Computational Linguistics*. Association for Computational Linguistics. (1995).

[8] Zhu, X.: Semi-supervised learning literature survey. *Computer Science*. University of Wisconsin-Madison, 2(3), 4 (2006).

[9] Mandarin VerbNet online: http://verbnet.lt.cityu.edu.hk/

[10] Stevenson, M., Yorick W.: Word sense disambiguation. The Oxford Handbook of Comp. *Linguistics:* 249–265 (2003).

[11] Mihalcea, R.: Word sense disambiguation. Encyclopedia of Machine Learning. *Springer* US, 1027–1030 (2011)

[12] Chan, Y.S., Hwee, T.N., David, C.: Word sense disambiguation improves statistical machine translation. *Annual Meeting-Association for Computational Linguistics*. 45(1) (2007).

[13] Purandare, A., Pedersen, T.: Word sense discrimination by clustering contexts in vector and similarity spaces. In: *Proceedings of the eighth conference on computational natural language learning* (CoNLL-2004) at HLT-NAACL (2004).

[14] Karov, Y., Shimon, E.: Similarity-based word sense disambiguation. *Computational linguistics*.24(1), 41–59.

[15] Liu, M.C.: A frame-based morpho-constructional approach to verbal semantics (框架为本，构式为用-基于语料库实证的汉语动词语义分析与分类). In: *Empirical and Corpus Linguistic Frontiers,* eds. by Chunyu Kit and Meichun Liu. Beijing: China Social Sciences Press. (2018).

[16] Liu, M.C.: Emotion in lexicon and gram-mar: lexical-constructional interface of Mandarin emotional

predicates. *Lingua Sinica* 2(4) (ERIH Plus, Springer Open Access). (2016).
[17] Liu, M.C., Chang, C.W.: A lexical-constructional approach to verbal semantics: the case of Mandarin 'hang' verbs. *International Journal of Knowledge and Language Processing* 6(4): 1–20. (2015).
[18] Liu, M.C., Chiang, T.Y.: The construction of Mandarin VerbNet: *a frame-based study of statement verbs. Language and Linguistics* 9(2), 239–270 (SSCI and AHCI index). (2008).
[19] Fillmore, C.J.: Frame semantics. Linguistics in the morning calm. (1982).
[20] Fillmore, C.J.: Frames and the semantics of understanding. *Quaderni di semantica*, 6(2), 222-254. (1985).
[21] Goldberg, A.E.: Constructions: A construction grammar approach to argument structure. University of Chicago Press. (1995).
[22] Langacker, R.W.: Foundations of cognitive grammar: Theoretical prerequisites. Stanford university press. (1987).
[23] Levin, B.: English verb classes and alternations: A preliminary investigation. University of Chicago press. (1993).
[24] Talmy, L.: Lexicalization patterns: Semantic structure in lexical forms. *Language typology and syntactic description,* 3(99), 36–149. (1985).
[25] Van, V., Robert, D.: Exploring the syntax-semantics interface. Cambridge University Press. (2005).
[26] Witten, I.H., Frank, E., Hall, M.A., Pal, C.J.: Data Mining: Practical Machine Learning tools and techniques. Morgan Kaufmann (2016).
[27] Pycharm Community: https://www.jetbrains.com/pycharm/
[28] Stanford Parser: http://nlp.stanford.edu:8080/parser/index.jsp
[29] Weka main webpage: https://www.cs.waikato.ac.nz/ml/weka/
[30] Li, Y.H., Jain, A.K.: Classification of text documents. *The Computer Journal*, 41(8),537–546. (1998).
[31] Steinwart, I., Christmann, A.: Support vector machines. *Springer Science and Business Media*. (2008).
[32] Platt, J.C.: 12 fast training of support vector machines using sequential minimal optimization. *Advances in kernel methods*, 185–208. (1999).
[33] Joachims, T.: Text categorization with support vector machines: Learning with many relevant features. In: European conference on Machine Learning (pp. 137–142). *Springer*, Berlin, Heidelberg. (1998, April).
[34] Manning, C.D., Schütze, H.: Foundations of statistical natural language processing. MIT press. (1999).