Filtered collocations as features in verbal polysemy disambiguation

A case study of the Chinese verb *kao* 'bake'

Yu-Yun Chang and Shu-Kai Hsieh
National Taiwan University

In Generative Lexicon Theory (GLT) (Pustejovsky 1995), co-composition is one of the generative devices proposed to explain the cases of verbal polysemous behavior where more than one function application is allowed. The English baking verbs were used as examples to illustrate how their arguments co-specify the verb with *qualia unification*. Some studies (Blutner 2002; Carston 2002; Falkum 2007) stated that the information of pragmatics and world knowledge need to be considered as well. Therefore, this study would like to examine whether GLT could be practiced in a real-world Natural Language Processing (NLP) application using collocations. We have conducted a fine-grained logical polysemy disambiguation task, taking the open-sourced Leiden Weibo Corpus as resource and computing with Support Vector Machine (SVM) classifier. Within the classifier, we have taken collocated verbs under GLT as main features. In addition, measure words and syntactic patterns are extracted as additional features for comparison. Our study investigates the logical polysemy of the Chinese verb *kao* 'bake'. We find that GLT could help in identifying logically polysemous cases; additional features would help the classifier achieve a higher performance.

Keywords: Generative Lexicon Theory, co-composition, baking verb, logical polysemy, collocation

1. Introduction

Word sense disambiguation (WSD) remains a tough issue waiting to be dealt with in Natural Language Processing (NLP). Most traditional WSD tasks using a lexico-syntactic analysis approach have failed to observe words beyond the surface level. This implies that in order to understand a word, information in all aspects need to be considered, whether syntactically, semantically, or cognitively. Generally, most linguists would assume that language meaning is compositional, but studies

Language and Linguistics 19:1 (2018), **61–79**. DOI 10.1075/lali.00003.cha
ISSN 1606-822X / E-ISSN 2309-5067

indicate that not every language phenomenon can be covered (Partee 1992; Kamp & Partee 1995). Hence, in order to explain linguistic complexity and train a much better machine learning model, various mechanisms have been proposed to perform fine-grained WSD tasks, of which Generative Lexicon Theory (GLT) is one. This paper focuses on applying GLT to Chinese examples and proposes a WSD model based on linguistic knowledge.

Recent work has tried to use various knowledge-based sources to support machine-readable resources (such as thesaurus, wordnet, and ontology), yet a knowledge-acquisition bottleneck has been encountered (Gale et al. 1993). For distinguishing senses in fine-grained granularity, many researchers would propose using a sense-enumerative approach (Wierzbicka 1996) within a sense inventory. Wierzbicka's (1996) Natural Semantic Metalanguage (NSM) proposed that various senses could be produced from a core meaning. However, it was noted that NSM only took actual words rather than conceptually abstract words for exhaustive listing (Goddard 1998). That is to say, an NSM approach would still be insufficient for treating lexical polysemy in a regular and proper way. Therefore, when it comes to the issue of motivating or organizing senses, instead of an unlimited enumerating of senses, a generative approach (Pustejovsky 1991; 1995) is suggested. The generative approach argues that senses are generated from rules, which infers that regularities could be structured.

In GLT, the co-composition operation is one of the generative devices that permit us to explain the polymorphic nature of verbs. That is, under a compositional interpretation, arguments of verbs could shift the meaning in some verbal alternations. Although, this poses difficulties to WSD tasks in contextualizing underlying senses, i.e. putting semantic weights on non-functor elements, to give rise to a derivative sense, GLT has proposed a plausible semantic type of framework (e.g. qualia structures) to be used in NLP.

Still some concerns regarding GLT have been addressed (Fodor & Lepore 1998; Blutner 2002; Falkum 2007). Falkum (2007) stated that GLT had some issues that needed to be considered and the framework of relevance theory (Sperber & Wilson 1995; Carston 2002) was needed. The example {Mary *baked* the pizza.} provided by Falkum (2007) illustrates two possible interpretations, which are "pizza is an artifact and Mary baked it (creative reading)" and "Mary simply heats up a frozen pizza (non-creative reading)". This kind of circumstance could not be covered properly by GLT. Therefore, these studies believe that while dealing with word meaning, pragmatics and world knowledge (i.e. contextual information) must be taken into consideration. Despite the fact that a range of objections have been leveled against GLT, this paper would like to investigate how GLT could be applied to NLP, and add additional features for training to examine whether the results could counter the above objections. To begin with, we start by analyzing the logical polysemy of *change of state* and *creation*.

Regarding co-composition in GLT, Pustejovsky (1995) took the English verb *bake* to explore the logical polysemy involving *change of state* and *creation* senses. When looking up the definitions of *bake* in Wordnet[1] for related senses and given examples, two listed senses are found related to *change of state* and *creation*. It must be pointed out, however, that the glosses provided for the two Wordnet senses do not clearly convey the contrast of *change of state* and *creation* senses. Although the examples given ("bake the potatoes" and "bake a cake") are cases embedded with the *change of state* and *creation* senses, the glosses of the two Wordnet senses do not reveal the distinction, e.g. "cook and make edible by putting in a hot oven" and "prepare with dry heat in an oven". This further implies that in order to individuate various senses of a lexical item, the embedded linguistic cues within examples need to be taken into account.

Since GLT typically uses English examples to illustrate the deep-level information of a lexical item, this study would also like to start by exploring whether *change of state* and *creation* senses, proposed by co-composition theory, could be applied to Chinese data. As Pustejovsky (1995) simply took the English verb *bake* as an illustrative example, we cleverly choose for this paper to apply GLT to the Chinese verb 烤 *kao* 'bake'.

As it is still difficult to fully apply an automatic WSD model, some manual tinkering has proven necessary.

Since collocations represent parts of context and could be easily extracted from various sources of tools or online services, we have attached linguistic-based features and filtering rules under the notion of GLT and other additional information to help optimize the WSD model. Therefore, this paper aims to examine whether GLT could be applied to real world NLP application, and tries to build a semi-automatic approach from the theory towards a deep-level linguistic WSD model. Within this paper, the model would be trained using SVM (i.e. Support Vector Machine, a machine learning approach) with features (including collocated verbs, measure words, and syntactic patterns) extracted from the open-sourced Leiden Weibo Corpus (van Esch 2012).[2] All the computational techniques are implemented using R built-in packages. Via this study, using a Chinese baking verb as an example and with features retrieved from collocations, we try to investigate how GLT could be employed in Chinese WSD tasks and analyze what other linguistic strategies could be further added in.

1. http://wordnetweb.princeton.edu/perl/webwn

2. Leiden Weibo Corpus collects messages from China's most popular micro-blogging platform, Sina Weibo. The corpus is already segmented and open-access: http://lwc.daanvanesch.nl/

2. Co-composition and qualia structure in GLT

As regard to lexical composition, although many semantic models agree that words have simple denotations, various perspectives on lexical composition have evolved. Some formal models would argue that composition approaches are truth-value denotation and computed with logical inferences; while in GLT, it is the semantic transformations (including type coercion, selective binding, and co-composition) of words' denotations that shift from one to another to form new meanings. Among these semantic transformations, co-composition is concerned in this study.

Co-composition is a semantic operation proposed to explain cases that are logically polysemous, introduced by Pustejovsky (1995) (originally named as co-specification (Pustejovsky 1991)) to capture the words' meanings. The term "logical polysemy" used by Pustejovsky (1995) was originally adapted from regular polysemy (Apresjan 1973). In general, polysemy is considered as carrying multiple but related meanings in the same lexical form, identified as complementary polysemy (Weinreich 1964). Weinreich (1964) stated that complementary polysemy could be distinguished as two types, which are lexical category preserving and lexical category changing. Examples of these two types are provided (Pustejovsky 1995):

- lexical category changing
  1. If the store is open, check the price of coffee.
  2. Zac tried to open his mouth for the dentist.
- lexical category preserving
  1. Mary painted the door.
  2. Mary walked through the door.

Following Weinreich (1964)'s sense distinctions, Pustejovsky (1995) defined logical polysemy as a complementary polysemy without changes in lexical category.

Although sense ambiguities could be enumerated (referred to as SEL (Sense Enumeration Lexicon); Pustejovsky 1995) as done in most dictionaries, this strategy is inadequate to address the nature of lexical knowledge. In order to maintain every aspect of compositionality, a systematic model, generative lexicon, has been proposed. Therefore, to better cover the compositionality of logical polysemy, an attempt at sense enumeration from a dictionary would not be recommended.

People might not need to disambiguate senses of logical polysemies in daily communication; while in NLP, identifying different senses of logical polysemies would be essential. Examples (1) and (2) are adapted from Pustejovsky (2005), and illustrated in the following.

(1) The rain started during the concert.

(2) The concert was confusing.

The word "concert" in Example (1) presents an event; while in Example (2), it describes the music played within the concert. More examples are shown in the following.

Example (3) and (4) are retrieved from Pustejovsky (1995).

(3) The lamb is running in the field.

(4) John ate lamb for breakfast.

The word *lamb* has two meanings in the above examples. In Example (3), *lamb* indicates a living animal; whereas in Example (4) it has the meaning 'meat'.

As could be seen from the above examples, there are different senses of *concert* and *lamb*. However, when looking up one of the words within dictionaries, the delicate senses of logical polysemies would not be addressed. On the other hand, it is not available for users to specify looking up one of the senses; for that, it would only return the results based on the dictionary definitions. Hence the application of logical polysemy under a framework is necessary in NLP.

It is observed that logical polysemy could be found in verbs, nouns, and adjectives (Pustejovsky 1995), but it is worth noting that not every word is logically polysemous. Atkins et al. (1988) illustrated a verbal logical polysemy case, *bake*, carrying two meanings: a sense of *change of state* and a sense of *creation*. When the verb *bake* is followed by a noun, the selection of senses would be produced via a co-composition mechanism. In addition, qualia structure could be applied to better specify a word's meaning. Qualia structure (Pustejovsky 1995), adapted from the modes of explanation by Aristotle, describes four main essential factors (*constitutive*, *formal*, *telic*, and *agentive*) to drive and capture the interpretation of an object as well as a relation (Moravcsik 1975).

Under the notion of co-composition and qualia structure, the verb *bake* itself is not polysemous but the nouns that followed have derived other meanings through *constitutive* and *agentive* roles. This can be examined from Example (51) provided by Pustejovsky (1995) (e.g. *bake a potato* and *bake a cake*). It is usually found that if the *constitutive* quale of a noun is an individual natural kind (e.g. default argument) such as *potato*, the selected sense for this logical polysemy would be *change of state* sense; whereas, if the *constitutive* quale is an artifact like *cake* which is composed of various components, the sense would turn out to be *creation* sense. Additionally, the *agentive* quale carries the information of explaining how an object comes into being, if an object remains after the process of comes into being, the verb *bake* would be *change of state* sense; however, when an object changed into another object after the baking process, *creation* sense would be assigned to the verb. This kind of event type shifting in a noun, is what makes the verb *bake* polysemous.

The process of co-composition dealing with cases of logical polysemy is listed below proposed by Pustejovsky (1995):

1. The governing verb would apply to its followed noun;
2. The noun would then co-specify the verb;
3. A new sense of the verb would be derived resulting from an operation called *qualia unification*, where the *agentive* roles of the verb and its noun match each other; and the *formal* quale of the noun is also the *formal* role of the entire verb phrase (VP).

Henceforth, in order to capture the amounts of meta-information within nouns in this study, we would like to investigate into qualia structure in order to fetch language behaviors as linguistic features.

It is found that within a verb phrase, not every part of knowledge embedded in the noun would be used as an effective factor to involve in sense selection. Therefore, to better capture the whole picture (e.g. qualia structure) of the noun, as many relations with its collocated verbs should be collected as possible. That is, so as to grab the complete qualia structure of a noun automatically, we have considered extracting all the collocated verbs from the Weibo corpus. Although this study tries to practice GLT using qualia structure, taking only collocated verbs as features might not be sufficient; in other words, some important factors beyond the theory might be ignored as noted by (Blutner 2002; Carston 2002; Falkum 2007). Thus, we would like to seek more features that are potentially related to the verb *kao* 'bake' via a collocation approach as well. These potential features would be used as additional information of the collocated verbs to further demonstrate whether these would have an impact on the GLT application within this study.

As to additional features, since traditional theories on lexical structure focus on discussing how verbs could be related to syntactic forms (Davis & Koenig 2000; Jackendoff 2002; Levin & Rappaport Hovav 2005; Van Valin 2005), some syntactic constructions of the verb *kao* 'bake' are considered as well. In addition, by observing the extracted examples of *kao* 'bake', it is found that measure words could also help classifying senses, for example, the measure word *chuan* '(a) string of' would mostly lead to *change of state* sense; while measure word *lu* '(an) oven of' would be *creation* sense. Therefore, Chinese measure words are also included in this study. The additional features used in this study would be extracted via collocations as well.

In this paper, we are interested in demonstrating whether GLT could be applied practically, by taking Chinese baking verb *kao* 'bake' cases as examples. In addition, we constructed a linguistic-based classifier in a semi-automatic way, under co-composition theory and qualia structure, using collocated verbs as features, and syntactic constructions and measure words as additional features. Below sections would show steps in classifier training procedures, results, discussions and conclusions.

3. Methodology

This section presents the procedure from data collection, filtering nouns, feature extraction to SVM classifier training. Since this study is conducted to examine the application of GLT, the collocated verbs are main features produced under GLT; while other features (e.g. measure words and syntactic patterns) are taken as additional features for testifying whether these could help improve the performance. All of the features are searched and retrieved from collocations.

3.1 Data collection

Since microblogging platforms are now widespread and become prevalent in social communication, these sources provide up-to-date language usage and have been taken as corpora in recent studies for data analysis. The Leiden Weibo Corpus (van Esch 2012; a large amount of open-sourced Chinese data from the microblog Weibo) is used in this paper. By using this corpus, we can not only freely access the data without expensive crawling techniques, but also observe the current linguistic cues. Therefore, with the corpus prepared, posts containing *kao* 'bake' can be easily retrieved using R programming language. R, due to its efficiency in data processing, convenience in applying statistical models and powerfulness in plotting, a total of 5,846 segmented posts involving the verb *kao* 'bake' have been successfully extracted for the following data analysis and classifier training.

3.2 Filtering nouns as seeds

From the 5,846 retrieved posts, we have tried to list out all the nouns that follow the verb *kao* 'bake'. It is observed that a verbal phrase such as *kao yige piaoliangde dangao* 'bake a beautiful cake', the noun *dangao* 'cake' is not adjacent to the verb; therefore in order to have the noun list include as many noun types as possible, the window size 5 is set to the right of the verb. In addition, the observed verb phrases such as *kao ershi fenzhong* 'bake for 20 minutes' are not within discussion of logical polysemy and would be excluded from the list.

There are 209 nouns found in the list and manually tagged as *change of state* and *creation* senses. However, the amount of *creation* senses is relatively smaller than *change of state* senses, with 10 and 197 respectively. Before using these nouns as seeds for classification, some noises need to be specified or removed. Filtering rules for nouns are listed as below:

- Nouns with *creation* senses
  Among the 10 nouns tagged as *creation* senses, cases like *kao mianbao* 'bake a bread', which could be either *change of state* or *creation* sense, are considered as well. This kind of circumstances would be left to be identified by additional features. The 10 nouns with *creation* senses within the list are, *shaobing* 'sesame pancake', *pisa* 'pizza', *danjuan* 'egg roll', *buding* 'pudding', *dangao* 'cake', *binggan* 'cookie', *mianbao* 'bread', *danta* 'egg tart', *xiangbing* 'naan', and *bulei* 'burnt cream'.
- Nouns with *change of state* senses
  Although many *change of state* senses are found in the corpus, in order to balance the numbers between the two senses, we have selected 15 cases which are a little bit more than *creation* senses. Steps for filtering out 15 cases are listed below:
  1. Choose cases with frequencies situated in the region of 75% quantile. This step is to ensure that the chosen cases could have enough collocated verbs and measure words to be extracted.
  2. Randomly select 15 cases from step (1).
  3. Examine the 15 cases with three rules to avoid overfitting the classifier.
     - Metonymy relation – if cases such as *yangtui rou* 'a leg of mutton' and *yang rou* 'mutton' have a metonymy relation, only the hyponym *yang rou* 'mutton' would be selected.
     - Lexicalized – words that are lexicalized would not be involved. For example, *kaorou* 'barbecue' and *kaoyangrouchuan* 'kabob' although contain the verb *kao* 'bake', the verb phrases are already lexicalized into noun phrases.
     - Having more than one metonymy relation with others (within the 15 cases), if found cases like *wuhuarou* 'animal's belly', *zhurou* 'pork' and *niourou* 'beef', *wuhuarou* 'animal's belly' would be removed from the list for the reason that it could be a part of more than one animal.
  4. Delete N (a variable presenting a specific number) cases that do not fit the rules, and randomly select N candidates from the pool in step (1).
  5. Re-examine step (4) with rules specified in step (3) until the 15 nouns are independent of each other.

The final group of the 15 nouns in *change of state* senses are *digua* 'sweet potato', *chang* 'intestine', *niurou* 'beef', *tudou* 'potato', *yangrou* 'mutton', *zhu* 'pig', *haixian* 'seafood', *doufu* 'tofu', *jitui* 'chicken drumstick', *mianjin* 'gluten', *baozi* 'baozi', *ya* 'duck', *shucai* 'vegetable', *ziba* 'glutinous rice cake', and *xiangjiao* 'banana'.

3.3 Selection of collocated features: verbs, measure words and syntactic constructions

After filtering out 15 *change of state* senses and 10 *creation* senses of nouns, the next step is to use these words to find out its collocated verbs, measure words and syntactic patterns as features for constructing a classifier.

3.3.1 *Collocated verbs and measure words*

In order to catch the most direct relations between the 25 nouns, and its collocated verbs and measure words, we have first of all arranged to extract all of the above collocations that were formerly adjacent to the verb (which is the first former gram of the target gram (*n-1* gram) from the verb).

Among the collocated verbs, only verbs that are *telic* or *agentive* roles to the corresponded nouns are selected. Again, to avoid an overfitting problem during classification, two examination rules are applied:

- Metonymy relation – if there is found a hyponym versus a hypernym within selected verbs, for example *fanchao* 'stir-fry' and *chao* 'fry', only the hypernym would be chosen while the hyponym is removed.
- Near-synonym – if there are cases like *lengdong* 'freeze' and *dong* 'freeze', these would be seen as one case only.

However, based on the above two rules, not all the 25 nouns have related collocations in the corpus. Under this situation, two out of 10 nouns with *creation* senses, *xiangbing* 'stuffed naan' and *bulei* 'burnt cream', are deleted.

It is noted that while using collocated verbs and measure words as features, cases such as *reng dangao* 'throw a cake', *yi che dangao* 'a cart of cakes' and *yi che binggan* 'a cart of cookies' would occur. It appears that it would then be insufficient to use these features. Nevertheless, these features are extracted through the collocation approach, and cases illustrated above are few which would have little influence on the results in machine learning. Even if an amount of such cases are found, this would indicate the collocated *reng* 'throw' as an important factor of *dangao* 'cake', and *che* 'cart' for *dangao* 'cake' and *binggan* 'cookie', presenting the current language use from corpus.

Therefore, though it seems that the collocated verb *reng* 'throw' and measure word *che* 'cart' might not have a direct relationship to help disambiguate *dangao* 'cake' and *binggan* 'cookie', it is still possible that the collocation preference derived from the corpus would have an impact on machine learning. Hence, we let the classifier learn from the corpus and handle the retrieved collocations automatically.

3.3.2 *Collocated syntactic patterns*

Observed from the extracted posts, four types of syntactic patterns frequently collocated with the verb *kao* 'bake' have been identified based on the definitions and tagsets in CKIP by 詞庫小組, the Chinese Knowledge Information Processing Group (1993),[3] which are listed below with examples found in the corpus. For convenience of illustrating the four categories in the following sections, category names are given as "Pattern 1" through "Pattern 4".

- Pattern 1 – *kao* 'bake' with complements: within such constructions, the verb indicates movement, whereas complements present results or directions as the supplementary information of the adjacent verbs, such as *kao chulai* 'bake and turn into', and *kao shou* 'bake something from raw to well-done'.[4]
- Pattern 2 – *kao* 'bake' with the particle *de* 得 'DE': are usually followed by status descriptions; for example, *kao de hen chenggong* 'successfully baked' and *kao de tai shou* 'over baked'.
- Pattern 3 – *kao* 'bake' with the preposition *ba* 把 'BA': this construction is used for expressing the meaning 'take something to do another act', such as *ba dangao kao yi xia* 'take the cake and bake it'.
- Pattern 4 – *kao* 'bake' with the adverb *le* 了 'LE': the verb followed with the aspect marker *le* 'LE' takes nouns as objects, like *kao le yi ge dangao* 'bake a cake'.

Therefore, in this study, there would be twenty-three nouns as seeds, including eight nouns with *creation* senses and fifteen with *change of state* senses. In addition, a total of 134 features are selected, which contain 89 verbs, 41 measure words, and 4 patterns. These features would be used to perform an SVM classification and investigate what kind of features could serve as important factors.

3.4 Constructing a data frame with features for SVM classification

To form a data frame for computing SVM, the rows are seeds with nouns tagged as *change of state* and *creation* senses and columns are set to be all of the collocated features. A Pointwise Mutual Information (PMI) value would be applied to each cell to compute the associations between nouns and features. The PMI equation is shown in Equation (5).

3. http://ckipsvr.iis.sinica.edu.tw/

4. In this pattern, only the two examples listed could be found from the corpus. Other complements such as kao re 'heat something up', kao ruan 'bake and turn something into a soft texture' and so on, are not observed in this corpus.

$$PMI = \frac{P(X,Y)}{P(X) \times P(Y)} \quad (5)$$

After PMI computation, the cells within the data frame are given with scaled values in order to get the data weighted in a normal distributed way. As a result, it is common that scaled numbers would be presented in positive and negative values.

3.5 SVM classification

Since the SVM approach is one of the most prevalent and effective classification technique nowadays, we use it to investigate the interactions between the 23 nouns and 134 features. To proceed with SVM, the data frame is randomly divided into two groups, of which 70% of the nouns are for training (16 nouns) and the 30% remaining for testing (7 nouns).

- training data (7 *creation* senses and 9 *change of state* senses): *shaobing* 'sesame pancake', *pisa* 'pizza', *danjuan* 'egg roll', *buding* 'pudding', *dangao* 'cake', *binggan* 'cookie', *danta* 'egg tart', *digua* 'sweet potato', *niurou* 'beef', *tudou* 'potato', *yangrou* 'mutton', *zhu* 'pig', *jitui* 'chicken drumstick', *mianjin* 'gluten', *ziba* 'glutinous rice cake' and *xiangjiao* 'banana'.
- testing data (1 *creation* sense and 6 *change of state* senses): *mianbao* 'bread', *chang* 'intestine', *haixian* 'seafood', *shucai* 'vegetable', *doufu* 'tofu', *ya* 'duck' and *baozi* 'baozi'.

The R package *e1071* is used to tune the classifier and to train it with a 10-fold cross validation. Furthermore, the results from SVM is presented by F-score in Equation (6), which returns a weighted average of the precision and recall values and the score ranges from 0 (the worst) to 1 (the best, which indicates cases in testing model are all predicted and classified correctly).

$$F = 2 \times \frac{precision \times recall}{precision + recall} \quad (6)$$

Via the SVM approach, we would like to investigate the interactions between nouns and the collocated verbs, and would also like to examine whether the additional features, measure words and syntactic constructions, could have an influence on the application of GLT in this current issue. Figure 1 presents the F-scores of SVM results by applying the collocated verbs and the corresponded combinations with additional features. Considering that features might be overfitted, feature selection is applied to get a more specific domain among the features, before processing the SVM approach to this data frame.

For each category of feature combination, only the top twenty features ranked by feature selection are chosen.

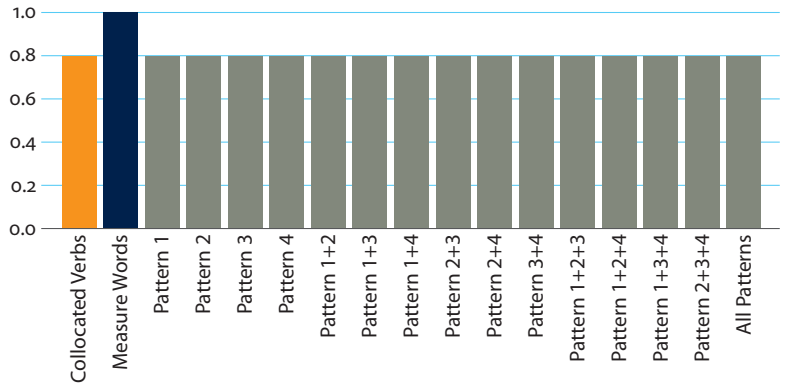

Figure 1. The F-score results of SVM by applying collocated verbs with additional features

From Figure 1 above, when using the collocated verbs as features, F-score returns 0.8 (the orange bar). However, when observing through the combinations with additional features, it is found that only when both the collocated verbs and measure words (one of the additional features) are taken into SVM training, the F-score would achieve 1.0 (the blue bar); whereas, for the rest of feature combination categories, the F-scores remain as 0.8 (the grey bars). So far, among the additional features, it seems that only measure words would serve as an important factor, while syntactic patterns present little effect.

In order to further confirm the impact of syntactic patterns, we take these features to directly train on the nouns without collocated verbs. Since there are only 4 syntactic patterns involved, feature selection would be ignored at this step. Figure 2 shows the F-scores of the SVM result by taking only Pattern 1 to Pattern 4 features and their various potential combinations.

As shown in Figure 2, the F-scores of Pattern 1 to Pattern 4, Pattern 1+2 and Pattern 2+3 are Not a Number (*NAN*), which refer us to the fact that there are no cases being categorized as carrying *creation* senses within testing data. However, it is found that some combinations of these patterns, such as Pattern 1+3, Pattern 1+4, and Pattern 2+4 would affect the classification results, and the F-scores are 0.67 (the orange bars). Among all of the combinations in Figure 2, the category of Pattern 3+4 has presented to be the most effective feature set with F-score equals to 1.0 (the blue bar).

As revealed in Figures 1 and 2, using the additional features, measure words, and Pattern 3+4 would succeed in classifying the *change of state* and *creation* senses under different approaches respectively. Therefore, in the following section, we would focus on analyzing and discussing these two feature sets.

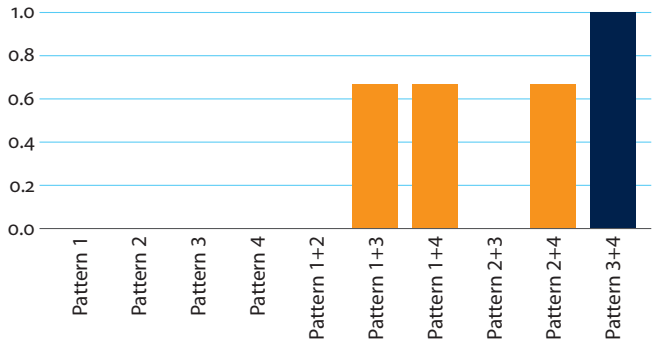

Figure 2. The F-scores of SVM by merely taking syntactic patterns as feature sets

4. Analysis and discussion

As the trained model could predict nouns accurately with either *change of state* or *creation* senses in two of the above feature sets, this section focuses on discussing the contributions of the two feature sets to the training data, which are the collocated verbs with measure words, and Pattern 3+4.

Since the frequencies in feature columns are scaled values, positive values within one column indicate that the feature would have a greater impact on the corresponded nouns, and vice versa for the negative values. Therefore, to figure out the interactions between these features and nouns, we could start by observing the scaled values. By picking out positive scaled values, the corresponded features and nouns are also extracted and further analyzed.

4.1 The collocated verbs with measure words

By taking the collocated verbs and measure words to train the SVM classifier, a total of 130 features are included. Among the 130 features, the top 20 features that are ranked by feature selection are chosen, which involve 10 verbs (*chao* 'stir fry', *shi* 'eat', *dong* 'freeze', *dingzhi* 'customized', *reng* 'throw', *zha* 'fry', *mai* 'sell', *yang* 'raise', *bo* 'peel' and *shao* 'cook') and 10 measure words (*dao* 'dish', *lu* 'stove', *bei* 'cup', *bao* 'packet', *che* 'car', *ke* 'gram', *jin* 'catty', *chuan* 'string', *xiang* 'box' and *ke* 'a (heart, corn grain, and so on)').

When reviewing the top 20 features with corresponded nouns, we find that these features could be further grouped into three classes, as shown in Table 1.

Table 1. The top 20 features that are grouped into three classes

Class	Description	Features	Corresponded nouns
Class 1	includes 11 features that would help to classify nouns into *change of state* senses	*chao* 'stir-fry', *dao* 'dish', *che* 'car', *ke* 'gram', *jin* 'catty', *chuan* 'string', *xiang* 'box', *yang* 'raise', *bo* 'peel', *ke* 'a (heart, corn, grain, etc.)', *shao* 'cook'	*digua* 'sweet potato', *niurou* 'beef', *tudou* 'potato', *yangrou* 'mutton', *zhu* 'pig', *jitui* 'chicken drumstick', *xiangjiao* 'banana', *mianjin* 'gluten'
Class 2	with 6 features that would assign nouns with *change of state* or *creation* senses	*shi* 'eat', *bei* 'cup', *bao* 'packet', *dong* 'freeze', *zha* 'fry', *mai* 'sell'	*digua* 'sweet potato', *niurou* 'beef', *tudou* 'potato', *yangrou* 'mutton', *xiangjiao* 'banana', *zhu* 'pig', *ziba* 'glutinous rice cake', *jitui* 'chicken drumstick', *pisa* 'pizza', *danjuan* 'egg roll', *binggan* 'cookie', *buding* 'pudding', *danta* 'egg tart', *shaobing* 'sesame pancake'
Class 3	contains 3 features that could identify nouns with *creation* senses	*lu* 'stove', *dingzhi* 'customized', *reng* 'throw'	*shaobing* 'sesame pancake', *dangao* 'cake'

4.1.1 *Features for nouns with* change of state *senses*

By observing the eleven features in Class 1, it is found that the corresponded cases share some characteristics and could be further analyzed through ontologies. For example, corresponded nouns like *niurou* 'beef', *yangrou* 'mutton', *zhu* 'pig' and *jitui* 'chicken drumstick' could be arranged under meat ontology; while *digua* 'sweet potato', *tudou* 'potato' and *xiangjiao* 'banana' would be considered instances of plant ontology. Generally speaking, instances that belong to the two ontologies would have a greater possibility of being recognized as embedded with *change of state* senses. In addition, it appears that the *constitutive* roles in most cases are individual natural kinds.

4.1.2 *Features for nouns with* creation *senses*

From the three feature classes, it is revealed that the number of features identifying *creation* senses are relatively smaller than those with *change of state* senses. As in Class 3, only three features are computed to have primary relations with *creation* senses, compared to Class 1 having eleven features for *change of state* senses.

Nonetheless, features in Class 2 could be seen as using to distinguish *creation* from *change of state* senses as well. Although features in Class 2 are found to have cases with either one of the two senses, it is observed that cases with *change of state* senses overlap with those in Class 1 and thus could be categorized into meat and

plant ontologies. While Class 2 cases with *creation* senses could not be categorized under Class 1, this indicates that features in Class 2 could be viewed as helping to identify a yes-no question – whether cases are also included in Class 1: if yes, cases would be assigned with *change of state* senses; if no, then could be referred to having *creation* senses.

Despite the fact that features in Class 2 are among the top twenty features, these features computed and chosen by machine might simply lead to the reason for the collocation frequencies being high within this corpus. Although the collocation frequencies might be corpus-dependent, it indicates a language usage distribution in that corpus. Despite this possible effect, we could note that features for recognizing *change of state* senses are more stabilized and fixed than *creation* senses.

4.1.3 *Others*

From Table 1, there are two cases tagged as *change of state* senses left unanalyzed, which are *mianjin* 'gluten' and *ziba* 'glutinous rice cake'.

It is found that the *constitutive* roles in the two cases involve more than one individual object. This contradicts the above finding which indicates the *constitutive* roles for *change of state* senses are individual natural kinds. However, there are no new beings generated after the two cases going through the baking process. According to GLT, this is what makes the cases carrying *change of state* senses; whereas if encyclopedic knowledge were involved, these conditions would be resolved. However, as regards the corpus-based machine learning process, it would be hard to grab the relations between objects and baking process due to the limitations on corpus quantity and quality. From the results in this study, the two cases though might not have the characteristics shared by most Class 1 cases, but their collocated frequencies with selected verbs and measure words are yet high enough to make the machine predict precisely.

4.2 Pattern 3+4

Although syntactic patterns did not present significant effects to the classification results when training with collocated verbs, it is found that when taking the feature set, Pattern 3+4, to directly train on nouns, the F-score would reach to 1.0. The corresponded nouns that are identified by this feature set are listed below:

- Pattern 3 – *ba* 'BA': *digua* 'sweet potato', *niurou* 'beef', *yangrou* 'mutton', and *jitui* 'chicken drumstick'
- Pattern 4 – *le* 'LE': *buding* 'pudding', *dangao* 'cake', *danta* 'egg tart' and *digua* 'sweet potato'

As could be seen from the above, under Pattern 3, most of the corresponded nouns are with *change of state* senses and could be categorized under meat ontology; while in Pattern 4, the nouns are mostly carrying *creation* senses. However, the case *digua* 'sweet potato' appears in both two features; in other words, the features could not identify whether *digua* 'sweet potato' is with *change of state* or *creation* sense. Examples of *digua* 'sweet potato' are presented below.

(7) 烤 了 幾 個 地瓜
kao le ji ge digua
bake LE some GE sweet potato
'bake some sweet potatoes'

(8) 用 烤箱 把 地瓜 烤 軟
yong kaoxiang ba digua kao ruan
use oven BA sweet potato bake soft
'use oven to bake sweet potatoes and make them softer'

The above two examples indicated that even with these features, the framework of GLT is still insufficient to depict the different senses of *digua* 'sweet potato' here and needs to include other linguistic knowledge (Falkum 2007). Despite the fact that Pattern 3+4 feature set might not be able to categorize the case *digua* 'sweet potato' correctly, these two features reveal some effects when disambiguating *change of state* and *creation* senses. In consequence, though syntactic patterns might not present obvious contributions due to limited amount of nouns when trained with collocated verbs, the interactions with nouns could not be neglected as well.

5. Conclusion

In this paper, we would like to examine and demonstrate how GLT could be applied practically using the Chinese verb *kao* 'bake' as an example with *change of state* and *creation* senses, and whether additional features would improve the performance. We present a semi-automatic linguistic-oriented way of training a Chinese WSD classifier under GLT. Though the preprocessing requires some manual work, we have listed systematic filtering rules that are machine-readable. By applying linguistically filtered collocations as features (including collocated features, measure words, and four syntactic features), via proposed steps for feature selection and training, the classifier has performed well and achieved 1.0 in F-score through some of the feature sets.

Among all of the feature sets trained with collocated verbs, the use of measure words presents having the best performance; while syntactic patterns do not reveal evident impacts. However, it is discovered that the combined feature set

Pattern 3+4 would provide certain contributions to nouns when trained without collocated verbs. This indicates that although the contributions of Pattern 3+4 are much less than measure words when trained with collocated verbs, the reason might lead to having a limited number of cases in this study. Accordingly, the effects to nouns could not be neglected. In addition, while analyzing the interactions between Pattern 3+4 and their corresponded nouns, it reveals that *digua* 'sweet potato' could not be successfully disambiguated via these two patterns. Therefore, there is still a need of replying on other syntactic features, such as Pattern 1 and Pattern 2. Even though Pattern 1 and Pattern 2 did not show significant effects during this classification, this did not imply that the two features would be worthless.

While discussing the feature sets of collocated verbs and measure words, it could be observed that features for *change of state* senses share some characteristics, which most of their cases could be grouped and sorted out under certain ontologies (e.g. plant ontology and meat ontology). It is worth noting that cases like *ziba* 'glutinous rice cake' and *mianjin* 'gluten' though tagged as *change of state* senses, could not be classified by simply applying ontologies. Henceforth, a fine-grained systematic semantic model such GLT needs to be included as well. Despite the fact that features for *creation* senses might not be as stabilized and patterned as *change of state* senses, via investigating the cases, it is not hard to find cases in this paper that could be taken as desserts which might leave a clue for further studies working on creation senses.[5]

In general, this study has worked well in disambiguating *change of state* and *creation* senses with some chosen features based on GLT and encyclopedic knowledge. By investigating the Chinese verb *kao* 'bake', using various types of collocations as features to train a fine-grained classifier is feasible. It is noted that F-score achieved 0.8 when merely trained on collocated verbs (selected based on GLT), which implied the contributions of GLT still existed. Correspondedly, the perspectives pointed out by Falkum (2007), Blutner (2002), and Carston (2002) should be valued in that additional features improved the performances.

During the procedure of feature selection, although some features might happen to be selected due to their high collocation frequencies instead of the weighted importance to SVM classifier, collocations could still indicate the effect of linked relations between target words and their collocations which would present up-to-date language usage within a corpus.

---

5. Noted that in this study though all the cases with *creation* senses are considered as different kinds of desserts, cases with *change of state* senses (e.g. ziba 'glutinous rice cake') could be desserts as well. Henceforth, using dessert ontology for disambiguating *change of state* and *creation* senses in future work needs further consideration.

Through this research examining how GLT could be applied practically using a Chinese baking verb as an example, we also have conducted a fine-grained sense disambiguation training procedure during this exploration. Based on this work, we could seek more Chinese logical polysemous cases in the future by applying ontologies, various collocation types, and pragmatic information, and adjust the proposed training procedure to organize a better structured and systematic approach that could better cover different cases.

Acknowledgements

Thanks go to the anonymous reviewers for providing valuable comments and suggestions, which are truly helpful and constructive to this work.

Abbreviations

GLT	Generative Lexicon Theory
N	a variable presenting a specific number
NAN	Not a Number, representing an undefined or unrepresentable value
n-1 gram	the first former gram of the target gram
NLP	Natural Language Processing
NSM	Natural Semantic Metalanguage
PMI	Pointwise Mutual Information
SEL	Sense Enumeration Lexicon
SVM	Support Vector Machine
VP	verb phrase
WSD	word sense disambiguation

References

Apresjan, Jurij D. 1973. Regular polysemy. *Linguistics* 142. 5–32.

Atkins, Beryl T. & Kegl, Judy & Levin, Beth. 1988. Anatomy of a verb entry: From linguistic theory to lexicographic practice. *International Journal of Lexicography* 1. 84–126. doi:10.1093/ijl/1.2.84

Blutner, Reinhard. 2002. Lexical semantics and pragmatics. *Linguistische Berichte* 10. 27–58.

Carston, Robyn. 2002. *Thoughts and utterances: The pragmatics of explicit communication*. Oxford: Blackwell. doi:10.1002/9780470754603

Chinese Knowledge Information Processing Group. 1993. *Technical report no. 93–05: Zhongwen cilei fenxi (sanban)* [The analysis of Chinese parts of speech (the third version)]. Taipei: Institute of Information Science, Academia Sinica.

Davis, Anthony R. & Koenig, Jean-Pierre. 2000. Linking as constraints on word classes in a hierarchical lexicon. *Language* 76(1). 56–91. doi:10.1353/lan.2000.0068

Falkum, Ingrid Lossius. 2007. Generativity, relevance and the problem of polysemy. *UCL Working Papers in Linguistics* 19. 205–234.
Fodor, Jerry A. & Lepore, Ernie. 1998. The emptiness of the lexicon: Reflections on James Pustejovsky's 'The generative lexicon'. *Linguistic Inquiry* 29(2). 269–288. doi:10.1162/002438998553743
Gale, William A. & Church, Kenneth W. & Yarowsky, David. 1993. A method for disambiguating word senses in a large corpus. *Computers and the Humanities* 26. 415–439. doi:10.1007/BF00136984
Goddard, Cliff. 1998. *Semantic analysis: A practical introduction*. Oxford: Oxford University Press.
Jackendoff, Ray. 2002. *Foundations of language: Brain, meaning, grammar, evolution*. Oxford: Oxford University Press. doi:10.1093/acprof:oso/9780198270126.001.0001
Kamp, Hans & Partee, Barbara. 1995. Prototype theory and compositionality. *Cognition* 57(2). 129–191. doi:10.1016/0010-0277(94)00659-9
Levin, Beth & Rappaport Hovav, Malka. 2005. *Argument realization*. Cambridge: Cambridge University Press. doi:10.1017/CBO9780511610479
Moravcsik, Julius M. 1975. Aitia as generative factor in Aristotle's philosophy. *Dialogue* 14. 622–636. doi:10.1017/S001221730002655X
Partee, Barbara H. 1992. Syntactic categories and semantic type. In Rosner, Michael & Johnson, Roderick (eds.), *Computational linguistics and formal semantics*, 97–126. Cambridge: Cambridge University Press. doi:10.1017/CBO9780511611803.004
Princeton University: About WordNet. 2010–2016. *WordNet*. (http://wordnet.princeton.edu) (Accessed 2014-09-01.)
Pustejovsky, James. 1991. The generative lexicon. *Computational Linguistics* 17(4). 409–441.
Pustejovsky, James. 1995. *The generative lexicon*. Cambridge: The MIT Press.
Pustejovsky, James. 2005. A survey of dot objects. Waltham: Brandeis University. (Manuscript.)
Sperber, Dan & Wilson, Deirdre. 1995. *Relevance: Communication and cognition*. 2nd edn. Oxford: Blackwell.
van Esch, Daan. 2012. *Leiden Weibo Corpus*. (http://lwc.daanvanesch.nl/index.php) (Accessed 2014-09-03.)
Van Valin Jr., Robert D. 2005. *Exploring the syntax-semantics interface*. Cambridge: Cambridge University Press. doi:10.1017/CBO9780511610578
Weinreich, Uriel. 1964. Webster's third: A critique of its semantics. *International Journal of American Linguistics* 30. 405–409. doi:10.1086/464799
Wierzbicka, Anna. 1996. *Semantics: Primes and universals*. Oxford: Oxford University Press.

Authors' addresses

Yu-Yun Chang (corresponding author)
Graduate Institute of Linguistics
National Taiwan University
No. 1, Sec. 4, Roosevelt Rd.
Taipei 10617
Taiwan
yuyun.unita@gmail.com

Publication history
Date received: 5 May 2015
Date accepted: 25 November 2016