Supervised Learning Algorithms Evaluation on Recognizing Semantic Types of Spanish Verb-Noun Collocations

Alexander Gelbukh
Olga Kolesnikova

Computación y Sistemas, 2012, to appear.

Abstract. The meaning of such verb-noun collocations as *the wind blows*, *time flies*, *the day passes by* can be generalized as '`exist` what is designated by the noun'. Likewise, the meaning of *make a decision*, *provide support*, *write a letter* can be generalized as '`make` what is designated by the noun'. These generalizations represent the meaning of certain groups of verb-noun collocations and may be used as semantic annotation. Collocations tagged with generalized meaning will be a valuable lexical resource for natural language applications like text analysis, text generation, machine translation, computer assisted language learning, etc. Our objective is to evaluate the performance of some existing machine learning methods which represent four learning approaches, namely Bayesian classification, rule derivation, decision tree construction, and nearest neighbor learning on the task of annotating collocations with the meanings `do`, `make`, `begin`, `continue`, `exist`, `act accordingly`, and `undergo` using hyperonyms of collocational elements as features. Experiments were carried out on a training set of Spanish verb-noun collocations applying 10-fold cross-validation technique. The obtained results show that methods based on rules and trees outperform Bayesian algorithms and achieve significant accuracy allowing them to be used in high quality semantic annotation.

Keywords: Collocations, Semantic Annotation, Supervised Machine Learning, Hyperonyms as Meaning Representation.

Evaluación de algoritmos de aprendizaje supervisado para reconocimiento de las clases semánticas de colocaciones verbo-sustantivo en español

Resumen. El significado de colocaciones de tipo verbo-sustantivo tales como *the wind blows*, el viento sopla, *time flies*, el tiempo vuela, *the day passes*, el día pasa, se puede generalizar y presentar con el patrón '`existe` lo que indica el sustantivo'. Análogamente, el significado de *make a decision*, tomar la decisión, *provide support*, proporcionar apoyo, *write a letter*, escribir una carta, se puede generalizar como '`hacer` lo que señala el sustantivo'. Estas generalizaciones representan el significado de ciertos grupos de colocaciones de tipo verbo-sustantivo y se pueden utilizar como anotación semántica. Las colocaciones etiquetadas con el significado generalizado, crearán un recurso léxico útil para aplicaciones de lenguaje natural tales como análisis y generación de textos, traducción automática y aprendizaje de idiomas asistido por computadora, etc. Nuestro objetivo es evaluar los algoritmos de aprendizaje de máquina supervisados como medio para etiquetar colocaciones de tipo verbo-sustantivo en español con los significados generalizados `do`, hacer (realizar), `make`, hacer (crear), `begin`, empezar, `continue`, continuar, `exist`, existir, `act accordingly`, actuar de acuerdo (con el sustantivo), y `undergo`, sufrir (por ejemplo, la consecuencia). Los algoritmos utilizados representan cuatro enfoques de aprendizaje, clasificación Bayesiana, derivación de reglas, árboles de decisión y el vecino más cercano. El funcionamiento de los algoritmos aplicados ha sido evaluado con el conjunto de entrenamiento usando la técnica de validación cruzada dividiendo el conjunto de entrenamiento en 10 muestras. Los resultados obtenidos muestran que los métodos de aprendizaje de máquina supervisados logran una precisión alta y se pueden utilizar para etiquetar colocaciones de tipo verbo-sustantivo con la información semántica representada por el significado generalizado.

Palabras Clave: Colocaciones, Anotación semántica, Aprendizaje de máquina supervisado, Hiperónimos como la representación del significado.

1 Introduction

Collocation is such a word combination in which one word, called **the base**, is used in its typical sense and the other word, called **the collocate**, is not used in its typical and well-predicted sense but acquires another meaning depending on the base. For example, in the collocation *to take a look*, *to take* does not mean 'to get into one's hands or into one's possession, power, or control' (sense 1 of *to take* in *Merriam-Webster Open Dictionary*, available online, see References), but 'to undertake and make, do, or perform' (sense 17 in the same dictionary). On the contrary, in free word combinations, words are used in their typical senses. Examples of free word combinations with *to take* are *to take a book*, *to take a cup*, *to take a flower*, etc.

Collocations present a challenge for natural language processing because the choice of collocates is not motivated semantically but depends on lexical preferences of their respective bases. One and the same meaning can be expressed by different words depending on the base. By saying *to take a look*, we mean 'to look, or to "perform" a look', but if we want to convey the semantics of "to perform" as applied to the noun *laugh*, we will say *to give a laugh*, but not **to take a laugh*.

Why does *a look* choose the verb *to take* and *a laugh* prefer *to make* for delivering the same idea? Linguists have not yet found the answer to this question. In spite of a lack of theoretical explanation, the knowledge of collocations, or restricted lexical co-occurrence, should be made available to language applications, because such knowledge is very important for text analysis and text generation.

One way to resolve this problem is to expand existing dictionaries including the meanings which words acquire when functioning as collocates. Since this requires a lot of manual work, it is time and money-consuming. Another problem caused by such an approach is that the number of senses per word in dictionaries will be increased making the hard disambiguation problem even harder.

Another way of dealing with collocations is to store them in special dictionaries of collocations only and use such dictionaries alongside with word dictionaries in language applications. The challenge is to create collocational dictionaries automatically. This involves automatic extraction of collocations and their semantic annotation. In this paper, we will give a brief discussion of the types of semantic annotation in Section 2 and propose a new type of semantic annotation which we call the generalized meaning in Section 3. The latter is related to the phenomenon termed lexical functions; a brief explanation of this topic will be presented as state-of-the-art in Section 4. Then in Sections 5-6 we study the performance of some algorithms representing four basic types of machine learning techniques, namely, Bayesian classification, rule induction, decision tree construction and the nearest neighbor method, on the task of annotating Spanish verb-noun collocation with seven generalized meanings. Finally, conclusions and future work will be outlined.

Some parts of this work were presented at the Mexican International Congress of Artificial Intelligence MICAI-2010, and published in its minutes.

2 Semantic Annotation

Semantic annotation of words in a corpus or in a wordlist is tagging words with names, attributes, comments, descriptions, or other labels which give additional information about the words. Semantic annotation resolves ambiguity of natural language by representing certain concepts in a formal language. The following kinds of semantic tags are commonly used in natural language applications:

1. Semantic or thematic roles. These tags generalize the semantics of verbal arguments with the purpose to map them to the syntactic frames. The verb expresses an action, event or state, and the verbal arguments are lexicalizations of the following concepts:

- agent (the doer of the action, event or the experiencer of the state),
- patient (the entity which experiences the consequences of the action/event/state), beneficiary (the entity that benefits from the action/event/state),
- goal (the location in the direction of which something moves),
- source (the location from which something moves),

- instrument (the means by which the action/event is carried out or the state is achieved),
- time (the location in time of the action/event/state),
- others.

Semantic roles are used in such projects as FrameNet, PropBank, UNL, SIL, EAGLES, among others.

2. Levin verb classes. Verb classification by Levin (1993) is based on the ability or inability of a verb to occur in pairs of meaning preserving syntactic frames (diathesis alternations) and on similar meanings. This classification is built on the assumption that syntactic frames reflect the underlying semantics.

As an example, let us consider two verbal classes: *break* verbs and *cut* verbs. Their common feature is that they both participate in the transitive and in the middle construction: *John broke the window*, *Glass breaks easily*, *John cut the bread*, *This loaf cuts easily*. The distinctive feature of *break* verbs is that only they can be used in the simple intransitive: *The window broke*, **The bread cut*. On the other side, only *cut* verbs can occur in the conative, *John cut at the dry bread with a knife*, **John broke at the window*.

Verb classes of Levin are implemented as a means of data organization in VerbNet, the largest on-line verb lexicon currently available for English.

3 Proposed Semantic Annotation: Generalized Meaning

The meaning of individual words can be described by definitions in conventional dictionaries for human usage like *Longman Dictionary of Contemporary English* (1995) or the *Merriam-Webster English Dictionary* (available online, see References). Often, most frequent words have many senses. For example, *Longman Dictionary of Contemporary English* gives 47 senses for the verb *to take*, 44 for *to make*, the number of senses for *to have* reaches 49, but *to play*looks very poor with only 10 senses! Combinations of verbs with prepositions, called phrasal verbs, like *to take after*, *to make over*, *to have on*, etc., are not counted as separate senses otherwise the number of senses would have grown tremendously!

Having taken a careful look at definitions of the previously given verbs, one will notice that these verbs have a particular meaning aspect in common and we are going to present this fact below. Note, that in this subsection we use word definitions from the *Longman Dictionary of Contemporary English* mentioned above. Therefore, when referring to the dictionary we mean the *Longman Dictionary of Contemporary English*.

Now we will show how some facets of meaning are repeated in verb definitions. We will do this by considering a few examples of polysemic verbs which have the meaning `do sth` (`sth` = something), among their other senses.

First, let us consider the verb *to take*. The dictionary gives the following definition of *to take* in the sense `do sth`: 'a word meaning to do something used with many different nouns to form a phrase that means: 'do the actions connected with the nouns': *take a walk* / *take a bath* / *take a breath* / *take a vacation*.'

The second example is the verb *to make*. In the dictionary, it also has the sense `do sth` followed by the comment: 'used with some nouns to mean that someone performs the action of the noun: *make a decision* / *mistake*.'

Thirdly, even the verb *to have* which is typically used in the sense 'possess', can acquire the meaning `do sth` in combination with some nouns. In this meaning, *to have* is described as 'a word meaning to do something, used in certain phrases: *have a look* / *walk* / *sleep* / *talk* / *thing* / *a holiday* / *bath* / *shower'*.

Lastly, let us consider the verb *to play*. One of its meanings given in the dictionary is 'to take part in a game or sport' like golf, chess, etc. Though the exact phrase `do sth` or the exact word `do` is not encountered in the definition of *to play*, we look for the definition of *to take part* in the dictionary and find: *to take part* is 'to do an activity, sport etc. with other people'. Therefore, it can be affirmed that *to play* also has `do` as one of its senses, because in the definition of *to play*, we can substitute *to take part* by 'to do an activity, sport etc. with other people'.

We will call the meaning `do sth`, or just `do`, the generalized meaning of the verbs *to take*, *to make*, *to have*, and *to play*, since `do` is used in the first, more general, part of verb definitions. Consider a few more `do` verbs accompanied by their dictionary definitions and example sentences:

to give somebody / *sth a smile* / *laugh* / *shout* / *push* (do something – to smile, laugh, shout etc.: *He gave me a quick smile and a hug*. | *Ooh, the baby just gave a kick*!),

to conduct a survey / *experiment* / *inquiry* etc. (to carry out a particular process, especially in order to get information or prove facts: The *company conducted a survey to find out local reaction to the leisure center*),

to carry sth out (to do something that needs to be organized and planned: *They are carrying out urgent repairs*. | *A survey is now being carried out nationwide*. | *It won't be an easy plan to carry out*),

to ask (*a question*) (to say or write something in order to get an answer, a solution, or information: *That kid's always asking awkward questions*),

to teach (to give lessons in a school, college, or university: *The guy's been teaching in France for 3 years now*).

Likewise, the generalized meanings `make`, `begin`, `continue`, `exist`, `act accordingly`, `undergo`, `cause` can be determined. `Exist` is a generalization of the meaning of verbs which confirm the fact of existence of an entity identified by a noun. Such verbs are used as predicates in utterances, and the corresponding nouns are grammatical subjects. `Act accordingly` means to act according to, or meet, the requirements of the entity expressed by the verb's direct object. `Undergo` is to be a patient of an action or event expressed by the direct object.

Below examples of all generalized meanings given above are given:

- `make`: *to create* (to make something exist that did not exist before: *Her behaviour was creating a lot of problems*), *to build* (to make something, especially a building or something large: *Are they going to build on this land?*), *to produce* (to make things to be sold: *Gas can be produced from coal*), *to write* (to produce a new book, poem, song etc.);
- `begin`: *to start* (to begin doing something: start learning German / work), *to enter* (to start working in a particular profession or organization: *Andrea is studying law as a preparation for entering politics*), *to introduce* (be the start of; if an event introduces a particular period or change, it is the beginning of it: *The death of Pericles in 429 BC introduced a darker period in Athenian history*), *to launch* (to start something, especially an official, public, or military activity that has been carefully planned: *launch a campaign* / *appeal* / *inquiry*), *to become* (to begin to be something: *He became King at the age of 17*);
- `continue`: *to keep* (to continue to have something and not lose it or get rid of it: *No, we're going to keep the house in Vermont and rent it out*), *to maintain* (to make something continue in the same way or at the same high standard as before: *Britain wants to maintain its position as a world power*), *to pursue* (to continue doing an activity or trying to achieve something over a long period of time: *Kristin pursued her acting career with great determination*), *to sustain* (to make something continue to exist over a period of time: *The teacher tried hard to sustain the children's interest*), *to run* (to continue to be officially able to be used for a particular period of time: *The contract runs for a year*).
- `exist`: *the possibility exists*, *time flies*, *the day passes by*, *a doubt arises*, *joy fills* (*somebody*), *the wind blows*, *an accident happens*, *the rain falls*;
- `act accordingly`: *to use a tool*, *to correct an error*, *to reach a level*, *to fulfill the obligation*, *to solve a problem*, *to utilize technology*, *to answer a question*, *to meet a standard*;
- `undergo`: *to get a benefit*, *to have an attack* (*of a disease*), *to receive treatment*, *to gain attention, to sit an exam*, *to face an accusation*, *to undergo an inspection*, *to take advice*.

It is a generally accepted fact that the meaning of an individual word depends on its context, i.e. the surrounding words in corpora. This fact is also true in the case of generalized meanings that we have determined. Verbs acquire these meanings when collocate with nouns belonging to a particular semantic group, for example, a group denoting actions. If verb-noun combinations are annotated with generalized meanings like do, make, begin, continue, exist, act accordingly, undergo, cause, etc., such annotation disambiguates both the verb and the noun. Word sense disambiguation is one of the most important and challenging tasks of natural language processing, and therefore semantic annotation of verb-noun combinations is a task of significant relevance. As it was also mentioned in Section 1, collocations tagged with semantic information (and the generalized meaning certainly is semantic information) may be a valuable lexical resource for natural language applications like text analysis, text generation, machine translation, computer assisted language learning, etc.

4 Related Work

4.1 Lexical Functions

It should be noted here that the concept of generalized meaning we propose in this word is close to the notion of lexical functions developed by the Meaning-Text Theory (Mel'čuk, 1974).

Lexical function is a mapping from one word (called **the keyword**, for example, decision) to another it collocates with in corpora (called **the lexical function value**). This mapping is further characterized by the meaning of semantically homogeneous groups of values and by typical syntactic patterns in which lexical function values are used with their respective keywords in texts. For the keyword *decision*, the lexical function $Oper_1$, meaning 'do, perform, carry out', gives the value *to make*. That is, to express the meaning 'do, or perform, a decision', one says in English *to make a decision*.

The formalism of lexical functions is intended to represent fixed word combinations, or collocations like *to make a decision*, *to give a lecture*, *to lend support*, etc. For more information on lexical functions, consult (Mel'čuk, 1996).

We do not apply the formalism of lexical functions as it is. Our purpose is to annotate verb-noun collocations with generalized meanings, and the meanings we have chosen are not exactly the meanings of lexical functions though have some resemblance to them. Another difference is that lexical functions describe collocations, but generalized meanings are present in collocations as well as in free word combinations. However, the research is made for collocations, not for free word combinations, and this focus is justified by the importance of collocations in natural language processing as explained in Section 1.1.

4.2 Automatic Tagging of Collocations with Lexical Functions

A few attempts have been made to annotate collocations with lexical functions automatically. One of the attempts is reported in (Wanner, 2004; Wanner et al., 2006) where semantic annotation of Spanish verb-noun collocations was viewed as a classification task. Classes were represented by nine lexical functions chosen for experimentation. These lexical functions had the meaning 'perform, experience, carry out something, 'cause the existence of something, 'begin to perform something, 'continue to perform something', etc.

Concerning linguistic data, (Wanner, 2004; Wanner et al., 2006) used two groups of Spanish verb-noun collocations. In the first group, the nouns belonged to the semantic field of emotions; in the second groups, the nouns were field-independent.

For classifying collocations according to lexical functions, the following supervised learning algorithms were applied: Nearest Neighbor technique, Naïve Bayesian network, Tree-Augmented Network Classification technique and a decision tree classification technique based on the ID3-algorithm.

As a source of information for building the training and test sets, the hyperonym hierarchy of the Spanish part of EuroWordNet was used (Vossen, 1998).

A hyperonym of a word A is a word B such that B is a kind of A. For example, *flower* is a generic concept for *rose*, *daisy*, *tulip*, *orchid*, so *flower* is hyperonym to each of those words. In its turn, hyperonym of *flower* is *plant*, and the hyperonym of *plant* is *living thing*, and hyperonym of *living thing* is *entity*. Thus, hyperonyms of a single word form a chain (*rose* → *flower* → *plant* → *living thing* → *entity*), and all words connected by the relation *kind-of*, or hyperonymy, form a tree.

Beside hyperonyms and synonyms, the hyperonym hierarchy in EuroWordNet also includes Base Concepts and Top Concepts. Base Concepts are labels of semantic fields like 'feeling', 'motion', 'possession'. Top Concepts, for example, 'Dynamic', 'Mental', 'Social', are words selected to characterize the Base Concepts.

The meaning of each verb-noun collocations was represented as a set of all synonyms, hyperonyms, Base Concepts and Top Concepts retrieved for the verb and for the noun. Each lexical function selected for the experiments had its own list of instances from which the prototypical instance was found. A candidate instance was assigned that lexical function whose prototype was the most similar to the instance. Similarity was measured using path length in hyperonym hierarchy.

The average F-measure of about 0.700 was achieved in these experiments. The best results for field-independent nouns were shown by ID3 algorithm (F-measure of 0.760) for the lexical function with the meaning 'cause (by the noun functioning in utterances as the verb's direct object) something to be experienced / carried out / performed' and by the Nearest Neighbor technique (F-measure of 0.737) for the lexical function with the meaning 'perform / experience / carry out something' (Wanner et al., 2006).

A more detailed analysis of the results achieved by (Wanner, 2004; Wanner et al., 2006) is given in Section 6, where state-of-the-art results are discussed together with the results obtained in our experiments.

5 Our Experimental Procedure

The objective of our work is to study performance of supervised machine learning methods on the task of annotating Spanish verb-noun collocations with generalized meanings. We have chosen methods which are characteristic of various commonly used approaches in machine learning: Bayesian classification, trees, rules, nearest neighbor technique, and kernel methods. We train the selected classifiers on a manually compiled corpus of verb-noun collocations tagged with generalized meanings and Spanish WordNet senses. After classification models having been built on the training data, the models are tested for annotating unseen data with the meanings. The tests are performed on the training set using 10-fold cross-validation technique.

5.1 Data

Verb-noun collocations for training sets were extracted automatically from *the Spanish Web Corpus* by the Sketch Engine (Kilgarriff et al., 2004) and ranked by frequency. This list contained 83, 982 pairs. From this list, we have taken the first one thousand pairs and processed them manually. The process of data preparation is schematized in Fig. 1. Now we present a detailed description of data preparation process, including a brief characteristic of resources used in that process.

The Spanish Web Corpus was the source of verb-noun pairs used in our experiments. It includes 116 900 060 tokens. It is compiled of texts found in the Internet. The texts are not limited to particular themes so the corpus represents the general Spanish lexis. The only peculiarity is that the texts are encountered on the World Wide Web and therefore reflect topics of public interest discussed there. *The Spanish Web Corpus* is available online, see References.

Verb-noun pairs were extracted automatically from the Spanish Web Corpus by means of the Sketch Engine (Kilgarriff et al., 2004), online software for automatic processing of corpora. Other tools can also be used for extracting verb-noun pairs, e.g. (Sidorov, 1996; Castro-Sánchez & Sidorov, 2010).

Then we removed all fallacious combinations extracted from the Spanish Web Corpus automatically due to parsing errors. Erroneous pairs included, for instance, past participles or infinitives instead of nouns, or contained symbols like --, «, © instead of words. The total number of erroneous pairs was 61, so after their removal the list contained 939 pairs.

Having removed erroneous pairs, we disambiguated each verb and noun, annotating them with word senses of the Spanish WordNet (Vossen, 1998; Spanish WordNet online). For some verb-noun pairs, relevant senses were not found in the above mentioned dictionary, and the number of such pairs was 39. For example, in the combination *dar cuenta*, 'to give account', the noun *cuenta* means 'razón, satisfacción de algo', reason, or satisfaction of something. This sense of *cuenta* is taken from *Diccionario de la Lengua Española* de Real Academia Española (2001) (The Spanish Language Dictionary). Unfortunately, this sense is absent in the Spanish WordNet so the expression *dar cuenta* was left without sense annotation. All combinations that could be not annotated with senses of the Spanish WordNet were removed from the list. After this step, 900 verb-noun pairs were left in the list.

Next, we have looked through the list of verb-noun combinations and annotated all relevant collocations with the meanings `do`, `make`, `begin`, `continue`, `exist`, `act accordingly`, and `undergo` manually.

As a result, the collected data included 266 collocations with the meaning `do`, the meaning `make` was represented by 109 collocations, 24 for `begin`, 16 for `continue`, also 16 collocations with the meaning `exist`, 60 collocations with the meaning `act accordingly`, the meaning `undergo` was encountered in 28 collocations. Thus, the total number of verb-noun collocations annotated with seven meanings was 519. The resting collocations were also annotated, but with other generalized meanings which we did not treat in this work.

All 519 collocations were included in the training set. Table 1 demonstrates examples of the data. The examples are given as they are encountered in the list compiled automatically, so the nouns are used without articles or quantifiers.

For machine learning methods to be applied, each data instance should be represented by a set of features characteristic for this instance. Hyperonyms were chosen as data features in our experiments. Therefore, the meaning of each noun and each verb was represented as a set of all hyperonyms of this noun or verb. Hyperonyms were extracted from the Spanish WordNet (Vossen, 1998; Spanish WordNet online). The meaning of a verb-noun collocation was thus represented as the union of the set of all hyperonyms of the verb and the set of all hyperonyms of the noun. Sets of hyperonyms also included both constituents of verb-noun collocations; collocational constituents were considered as zero-level hyperonyms.

It should be noted here, that the Spanish WordNet is structured the same way as the Princeton WordNet (Fellbaum, 1998). In the latter, nouns, verbs, adjectives, and adverbs are organized into synonym sets, or synsets, each representing one underlying lexical concept. When hyperonyms are extracted from the Spanish WordNet, what we actually obtain is synsets of hyperonyms. Each synset has its identification number, and every word in a synset is tagged with a sense number.

Table 1. Examples of verb-noun collocations

Generalized meaning	Collocations	
	Spanish	English lit. translation
`do`	*hacer justicia* *realizar actividad* *dar beso*	*do justice* *realize activity* *give kiss*
`make`	*hacer ruido* *establecer criterio* *encontrar solución*	*make noise* *establish criterion* *find solution*
`begin`	*iniciar proceso* *tomar iniciativa* *adoptar actitud*	*initialize process* *take initiative* *adopt attitude*
`continue`	*mantener control* *llevar vida* *seguir curso*	*maintain control* *lead life* *follow course*
`exist`	*relación existe* *diferencia existe* *año pasa* *sábado pasa*	relation exists difference exists year passes Saturday passes
`act accordingly`	*alcanzar meta* *conseguir objetivo* *cumplir requisito* *satisfacer demanda*	reach aim achieve objective fulfill requirement satisfy demand
`undergo`	*obtener resultado* *recibir ayuda* *recoger información* *sufrir daño*	obtain result receive help get information suffer damage

Let us consider an example. Suppose we want to build a meaning representation for the collocation *recibir ayuda*, to receive help. Since all collocations in the training set are labeled with the senses of their components, hyperonyms of *recibir*_1 and *ayuda*_1 must be looked for.

First, the synset containing *ayuda*_1 is retrieved: 00782440n *asistencia*_1 *ayuda*_1 (assistance, help) where 00782440n is the synset's identification number. There are two hyperonym synsets for the synset with *ayuda*_1: 00261466n *actividad*_1 (activity) and 0017487n *acto*_2 *acción*_6 (act, action). Therefore, the meaning representation of *ayuda*_1 is the set {*asistencia*_1 *ayuda*_1; *actividad*_1; *acto*_2 *acción*_6}.

Likewise the verb's meaning representation is constructed which is the set {*recibir*_1; *conseguir*_1 *tomar*_1 *sacar*_1 *obtener*_1}.

Lastly, the meaning representation of *recibir*_1 *ayuda*_1 is build and we get the set {*asistencia*_1 *ayuda*_1; *actividad*_1; *acto*_2 *acción*_6; *recibir*_1; *conseguir*_1 *tomar*_1 *sacar*_1 *obtener*_1}. In this set, each hyperonym synset is considered a feature.

5.2 Methodology

Two types of experiments were carried out. First, we studied the performance of diverse machine learning algorithms on the task of annotating Spanish verb-noun collocations with generalized meanings. The purpose was to identify algorithms which operated best. The task was viewed as binary classification; i.e., predicting if a particular collocation belongs to a given class or not. Each of seven generalized meaning was represented as a class variable with two possible values: 1 if a given collocation is of that class and 1 if it is not. Experiments were fulfilled on 42 algorithms using WEKA version 3-6-2 software (Witten & Frank, 2005; Hall et al., 2009; The University of Waikato, WEKA download):

Class `bayes`: AODE, AODEsr, BayesianLogisticRegression (BLR), BayesNet, HNB, NaiveBayes, NaiveBayesSimple, NaiveBayesUpdateable, WAODE.

Class `functions`: LibSVM, Logistic, RBFNetwork, SimpleLogistic, SMO, VotedPerceptron, Winnow.

Class `lazy`: IB1, IBk, KStar, LWL.

Class `rules`: ConjunctiveRule, DecisionTable, JRip, NNge, OneR, PART, Prism, Ridor, ZeroR.

Class `trees`: ADTree, BFTree, DecisionStump, FT, Id3, J48, J48graft, LADTree, RandomForest, RandomTree, REPTree, SimpleCart.

Secondly, the task of annotating Spanish verb-noun collocations with generalized meanings was viewed as *k*-class classification problem. Each meaning was seen a category, thus we had 7-class classification. To perform such classification, we chose a number of methods which may be called characteristic of various commonly used approaches in machine learning: Bayesian classification, rule induction, decision tree construction, the nearest neighbor technique, and kernel methods. For experimentation, the following classifiers implemented in WEKA have been chosen: NaiveBayes for Bayesian classification; PART, JRip, Prism; Ridor for rules; BFTree, SimpleCart, FT, REPTree for trees; IB1 for the nearest neighbor approach, and SMO for kernel techniques.

In both types of experiments, the performance of algorithms was evaluated on the training set using 10-fold cross-validation (Kohavi, 1995).

6 Results and Discussion

Table 2 and Table 3 present the results of the performance of algorithms chosen for the first type of experiments as explained in Section 5.2. Five best algorithms were identified for each of seven generalized meanings, in Tables 2-3 they are listed ranked by the values of F-measure. Average F-measure is given for each generalized meaning.

It is seen from Tables 2-3, six of seven best algorithms tested on the task of annotating Spanish verb-noun combinations with generalized meanings `do`, `make`, `begin`, `continue`, `exist`, `act accordingly`, and `undergo` belong to the category of rule-based techniques. The purpose of algorithms based on rules is to examine data and construct rules which are first-order conditional statements of the form presented below (remember that features in our data are hyperonyms):

If hyperonym$_1$ = *word*$_1$ **and** hyperonym$_2$ = *word*$_2$ ... **and** hyperonym$_n$ = *word*$_m$
then the collocation is of <generalized meaning>,

where <generalized meaning> is one of the seven meanings chosen for our experiments and mentioned above.

Rule-based methods acquire and use conceptual knowledge which is human-readable and easy to understand (Witten & Frank, 2005). Indeed, a concept consists of a number of features which are necessary and sufficient for description of an abstract idea. It appears that verb-noun collocations as specific linguistic data can be well distinguished by rules including hyperonym information. Since hyperonyms are words more generic than a given word, they are able to represent the generalized meaning. It should be remarked here that though hyperonyms and generalized meanings both depict the abstract meaning typical for a significant number of linguistic phenomena and in that sense both possess "generalized" nature, they are different in some important respects. However, this issue belongs more to the area of pure linguistics than to natural language engineering and we will not consider it here.

The best result is shown by the rule-based method PART with the value of F-measure of 0.877 for the meaning `do`. The second best result is achieved by Ridor (F-measure of 0.813 for the meaning `continue`). The third and the fourth places are held by Prism (0.781 for the meaning `act accordingly` and 0.757 for the meaning `begin`). The resting best algorithms are JRip reaching the value of F-measure of 0.716 for `make`, then again comes PART (0.706 for `undergo`), and the last best result of 0.696 is shown by BFTree for the meaning `exist`. Note, that BFTree is the only decision tree learning algorithm mentioned in this paragraph, the other six algorithms which were found to be best on annotating collocations with the generalized meanings are based on rules as it was said earlier in this section. Concerning decision tree algorithms, we will comment on them in the paragraph that follows.

Table 2. Best-performing learning algorithms for the meanings `do`, `make`, `begin`, `continue`

DO		**MAKE**	
rules.PART	**0.877**	rules.JRip	**0.716**
trees.SimpleCart	0.876	trees.SimpleCart	0.708
bayes.BLR	0.874	trees.LADTree	0.706
trees.BFTree	0.869	tress.REPTree	0.704
functions.SMO	0.864	trees.BFTree	0.699
Average	0.872	Average	0.707
BEGIN		**CONTINUE**	
rules.Prism	**0.757**	rules.Ridor	**0.813**
trees.FT	0.711	trees.REPTree	0.800
functions.SMO	0.683	lazy.LWL	0.800
rules.NNge	0.682	functionsLogistic	0.786
trees.Id3	0.667	rules.Prism	0.783
Average	0.700	Average	0.796

Table 3. Best-performing learning algorithms for the meanings `exist`, `act accordingly`, `undergo`

EXIST		**ACT ACCORDINGLY**		**UNDERGO**	
trees.BFTree	**0.696**	rules.Prism	**0.781**	rules.PART	**0.706**
trees.Id3	0.640	bayes.BLR	0.650	trees.J48	0.706
trees.J48	0.636	functions.SMO	0.627	trees.LADTree	0.667
lazy.LWL	0.632	trees.FT	0.598	rules.JRip	0.629
bayes.BLR	0.600	rules.NNge	0.593	trees.SimpleCart	0.625
Average	0.641	Average	0.650	Average	0.667

It is also seen from Tables 2-3 that the second best type of machine learning algorithms on the task on annotating verb-noun collocations with the generalized meanings is trees. Decision tree learning is a technique whose purpose is to identify a discrete-valued target function as precisely as possible, and the function is represented by a decision tree (Mitchell, 1997). Trees can also be put in the form of rules to make them more human readable.

Table 4 present the results of the second type of experiments as explained in Section 5.2. Table 4 presents the results of the performance of algorithms which showed to be best in the first type of experiments: PART, JRip, Prism, and Ridor representing rule induction algorithms, and BFTree. To obtain more evidence of operation of decision tree methods, we also experimented with the following tree-constructing algorithms: SimpleCart, FT, and REPTree. We also studied the performance of NaiveBayes (NB in Table 4), a classical probabilistic Bayesian classifier, as well as the performance of IB1, a basic nearest neighbor instance based learner using one nearest neighbor for classification, and SMO (Sequential Minimal Optimization), an implementation of support vector machine.

Table 4. Performance of some algorithms using 7-class approach

Algorithm	DO	MAKE	BEGIN	CON-TINUE	EXIST	ACT ACC.	UNDER-GO	Weighted average
rules.PART: 21 rules	0.894	0.783	0.524	0.774	0.903	0.685	0.643	0.812
rules.JRip: 26 rules	0.878	0.800	0.634	0.800	0.903	0.686	0.667	0.815
rules.Prism: 222 rules	0.896	0.840	0.647	0.720	0.889	0.744	0.682	0.841
rules.Ridor: 31 rules	0.888	0.709	0.667	0.774	0.710	0.576	0.618	0.780
trees.BFTree	0.908	0.814	0.605	0.800	0.875	0.672	0.667	0.830
trees.SimpleCart	0.915	0.798	0.605	0.800	0.875	0.672	0.656	0.829
trees.FT	0.915	**0.863**	0.714	**0.875**	0.903	**0.757**	**0.733**	**0.865**
trees.REPTree	0.893	0.746	0.632	0.759	0.759	0.529	0.677	0.788
bayes.NB	0.759	0.698	0.000	0.000	0.000	0.119	0.000	0.549
lazy.IB1	0.783	0.620	0.378	0.519	0.688	0.462	0.444	0.664
functions.SMO	**0.916**	0.843	**0.773**	0.839	**0.933**	0.739	0.714	0.861

The best result in Table 4 is shown by SMO. This algorithm was able to reach the F-measure of 0.916 for predicting the meaning `do`. It is also the best among methods indicated in Table 4 for the meanings `begin` and `exist`. SMO achieved the second best weighted average for all seven generalized meanings. As it was mentioned above, SMO is an implementation of support vector machine (Cortes & Vapnik, 1995), a non-probabilistic binary linear classifier. For a given instance of training data, it predicts which of two possible classes the instance belongs to. A support vector machine model is a representation of the examples as points in space, mapped so that the examples of the separate classes are divided by a clear gap. New examples are then mapped into that same space and predicted to belong to a category based on which side of the gap they fall on. SVM have been used successfully for many NLP tasks, for example, word sense disambiguation, part of speech tagging, language identification of names, text categorization, and others. As our results demonstrate, it is also effective for annotating collocations with the generalized meanings.

However, FT (Functional Tree), a generalization of multivariate trees able to explore multiple representation languages by using decision tests based on a combination of attributes (Gama, 2004), was more successful than SMO for predicting the meanings `make`, `continue`, `act accordingly`, and `undergo` for which FT acquired the highest value of F-measure.

For rule induction algorithms, Table 4 includes the number of rules generated by each technique. It appears that PART and JRip are more effective since they show high values of F-measure, 0.812 and 0.815, respectively, and generate quite a modest number of rules (21 and 26) compared with Prism which in spite of a higher F-measure of 0.841 generates as much as 222 rules.

NaiveBayes (Jiang, Cai & Wang, 2010) and IB1 (Aha & Kibler, 1991) showed a well-noted tendency to have low F-measure for all meanings in the experiments. The failure of NaiveBayes may be explained by the fact that statistical methods are limited by the assumption that all features in data are equally important in contributing to the decision of assigning a particular class to an example and also independent of one another. However, it is a rather simplistic view of data, because in many cases data features are not equally important or independent. The latter is certainly true for linguistic data, especially for such a language phenomenon as hyperonyms (remember, the meaning of collocations is represented by hyperonyms in our training sets). Hyperonyms in the Spanish WordNet form a hierarchic structure where every hyperonym has its ancestor, except for the most general hyperonyms at the top of the hierarchy, and daughter(s), except for most specific hyperonyms at the end of the hierarchy.

However, Naive Bayes is one of the most common algorithms used in natural language processing, it is effective in text classification (Eyheramendy et al., 2003), word sense disambiguation (Pedersen, 2000), information retrieval (Lewis, 1998). In spite of that, it could hardly distinguish the generalized meanings of collocations in our experiments. In the previous paragraph some reasons for this failure are suggested.

Low results of IB1 demonstrates that normalized Euclidean distance used in this technique to find the training instance closest to the given test instance does not approximate well the target classification function. Another reason of low performance can be the fact that if more than one instances have the same smallest distance to the test instance under examination, the first one found is used, which can be erroneous.

Since all best classifiers for predicting the generalized meaning are rule-based, we can suppose that semantics of collocations is better distinguished by rules than on the basis of probabilistic knowledge learned from the training data.

This may be explained by the fact that statistical methods are limited by the assumption that all features in data are equally important in contributing to the decision of assigning a particular class to an example and also independent of one another. However, it is a rather simplistic view of data, because in many cases data features are not equally important or independent. The latter is certainly true for linguistic data, especially for such a language phenomenon as hyperonyms (remember, the meaning of collocations is represented by hyperonyms in our training sets). Hyperonyms in the Spanish WordNet form a hierarchic structure where every hyperonym has its ancestor, except for the most general hyperonyms at the top of the hierarchy, and daughter(s), except for most specific hyperonyms at the end of the hierarchy.

Better performance of rule-based methods on predicting collocational semantics as it was proposed in our work, lead to an important issue which has a great impact on how natural language applications can be developed. Certainly, computer applications make use of linguistic knowledge, but this knowledge should be carefully selected and proved to be necessary and sufficient for the computer to analyze and generate texts in natural language effectively.

At present, there are two types of knowledge taken into account at building language applications, namely, the statistical knowledge and the symbolic one. According to these two types of information, two approaches to natural language processing have emerged. The first approach aims at building systems applying linguistic rules, which can be rather numerous and sophisticated. The second approach takes advantage of statistic information like word frequencies and distributions.

As it was observed above, rule-based methods outperformed statistical methods in distinguishing among the meanings of collocations in our experiments. It can be concluded that collocations are analyzed better by rules than by frequency counts; in other words, rules tell us more of what collocations are than frequency counts do and that knowledge in terms of rules is more informative than knowledge in terms of numbers.

Another conclusion that can be drawn from a better performance of rule-based methods concerns theoretical aspects of linguistics, in particular the definition of collocation. But since this article is oriented towards the computational side of computational linguistics, we leave this issue without further discussion. A more linguistically-oriented reader may consult (Wanner, 2004) for an overview of the dispute between two "camps": the adherents of the statistical approach to the definition of collocations (beginning with M.A.K. Halliday, see his definition of collocation in (Halliday, 1961) and those who claim that the semantic criterion is crucial for distinguishing collocations from all other word combinations as in (Mel'čuk, 1995).

The representation of collocational semantic content in the form of generalized meaning as explained in Section 3 is a new way to view lexical and semantic information that can be disclosed by analyzing word co-occurrences. Our attempt to annotate collocations with generalized meanings has been quite satisfactory.

The only research we can refer to when considering our results is (Wanner et al., 2006) because the concept of lexical function is similar to the generalized meaning proposed in this work. It was mentioned in Section 4.2 that while discussing our results, a more detailed presentation and analysis of state-of-the-art results in (Wanner, 2004) and (Wanner et al., 2006) would be given. Three points should be mentioned here. Firstly, we will compare state-of-the-art results for those lexical functions (explained in Section 4.1) whose semantics is closest to the generalized meanings used in our experiments. Secondly, in (Wanner, 2004) and (Wanner et al., 2006), the experiments were carried out on two sets of verb-noun collocations, as it was explained in Section 4.2. The first set included verbal collocations with emotion nouns, in the second set, the nouns were field-independent. Collocations in our corpus are field-independent, so we will compare only the results for field-independent collocations from (Wanner, 2004) and (Wanner *et al*., 2006) with our experimental results. Thirdly, we run the experiments on data other than in (Wanner, 2004) and (Wanner *et al*., 2006) and moreover, our data is annotated with the generalized meanings as described in Section 5.1 but not with lexical functions. Due to inequality of data sets, it is not fair to compare the results. However, we take the liberty to make such a comparison as a way of presenting the results achieved in the area of automatic semantic annotation.

Table 5. State-of-the-art results for automatic detection of lexical functions and performance of best algorithms on predicting the corresponding generalized meaning

LF / GM	# in W04	Result in W04, F	# in W06	Result in W06		# in our data set	Our result	
				F	Method		F	Method
$Oper_1$ / `do`	50	0.609	87	0.737	NB	266	0.877	rules.PART
$Oper_2$ / `undergo`	48	0.759	48	0.662	NN	28	0.706	rules.PART
$CausFunc_0$ / `make`	53	0.766	53	0.676	NN	109	0.716	rules.JRip
$Real_1$ / `act accord.`	52	0.741	52	0.500	NN	60	0.781	rules.Prism
Average		0.719		0.644			0.770	

Therefore, Table 5 presents some best results reported in (Wanner, 2004) and (Wanner *et al*., 2006), together with our results obtained in the experiments of the first type. In this table, LF stands for lexical function, GM stands for the generalized meaning, W04 and W06 are (Wanner, 2004) and (Wanner et al., 2006), respectively, F stands for F-measure, # signifies the number of instances for a given LF, NB is Naive Bayes, NN is the nearest neighbor algorithm. Results in W04 are demonstrated by the nearest neighbor method. The results in W06 are given in terms of precision and recall but here we present them as values of F-measure which is the harmonic mean of precision and recall. F-measure was computed by us to make the comparison with W04 and our results easier.

Another remark is important here. Data representation in our work is different than of (Wanner, 2004) and (Wanner et al., 2006). Section 4.2 explained that in order to make the meaning of collocations accessible to supervised classifiers, the collocations were represented in (Wanner, 2004) and (Wanner et al., 2006) as sets of hyperonyms, Base Concepts and Top Concepts. In our research, only hyperonyms were included in the data sets. However, the results of our experiments are better although such features as Base Concepts and Top Concepts were absent in our data representation. It seems that these features do not assist in distinguishing among generalized meanings. Nevertheless, for the meanings `undergo` and `make` state-of-the-art results are higher. Therefore, further research is necessary to determine the importance of Base Concepts and Top Concepts as features in distinguishing among generalized meanings.

7 Conclusions and Future Work

It has been demonstrated that it is feasible to apply machine learning methods for predicting the semantics of Spanish verb-noun collocations in the form of the generalized meaning proposed in this work. In particular, we studied the performance of learning algorithms on the task of assigning the generalized meanings `do`, `make`, `begin`, `continue`, `exist`, `act accordingly`, and `undergo` to a previously unseen verb-noun pair.

It has also been demonstrated that hyperonym information is sufficient for distinguishing among the generalized meanings. The best F-measure achieved in our experiments is 0.877 using the training set and 10-fold cross-validation technique. This result can be compared with results on the task of classification of collocations according to lexical functions since the concept of the lexical function is similar to the concept of generalized meaning. The highest F-measure achieved on classifying collocations using the taxonomy of lexical functions was 0.760. However, such a comparison is not fair due to differences in theoretical grounds and data.

In the future, we plan to test other semantic representations like word space models and explore the effect of other data features, such as WordNet glosses. We also plan to examine how the techniques of automatic selection of the best classification (Pranckeviciene, Somorjai & Tran, 2007; Escalante, Montes & Sucar, 2009) can be applied to the task of semantic annotation of collocations with generalized meanings. Another intention is to verify classification models on a test set and experiment with different ratios between the training set and the test set.

Acknowledgements. We are grateful to Adam Kilgarriff and Vojtěch Kovář for providing us a list of most frequent verb-noun pairs from the Spanish Web Corpus of the Sketch Engine, www.sketchengine.co.uk.

The work was done under partial support of Mexican Government: SNI, COFAA-IPN, PIFI-IPN, CONACYT grant 50206-H, and SIP-IPN grant 20100773.

A shorter version of the paper has already appeared in MICAI-2010.

References

Aha, D., Kibler, D. (1991). Instance-based learning algorithms. *Machine Learning*, 6, 37–66.

Castro-Sánchez, N. A. & Sidorov, G. (2010). Analysis of Definitions of Verbs in an Explanatory Dictionary for Automatic Extraction of Actants based on Detection of Patterns. *Lecture Notes in Computer Science* 6177, 233–239.

Cortes, C. & Vapnic, V. (1995). Support Vector Networks. *Machine Learning*, 20, 1–25.

Escalante, H. J., Montes, M., & Sucar, E. (2009). Particle swarm model selection. *Journal of Machine Learning Research*, 10, 405–440.

Eyheramendy, S., Lewis, D. & Madigan, D. (2003). On the Naive Bayes Model for Text Categorization. *Proceedings of the Ninth International Workshop on Artificial Intelligence and Statistics*.

Fellbaum, C. (Ed.) (1998). *WordNet: An Electronic Lexical Database*. Cambridge, MA: MIT Press.

Gama, J. (2004). Functional Trees. *Machine Learning*, 55(3), 219–250.

Halliday, M. A. K. (1961). Categories of the Theory of Grammar. *Word*, 17, 241–292.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P. & Witten I. H. (2009). The WEKA Data Mining Software: An Update. *SIGKDD Explorations*, 11(1).

Jiang, L., Cai, Z., & Wang, D. (2010). Improving naive Bayes for classification. *International Journal of Computers and Applications*, 32(3), 328–332.

Kilgarriff, A., Rychly, P., Smrz, P. & Tugwell, D. (2004). The Sketch Engine. *Proceedings of EURALEX*, 105–116.

Kohavi, R. (1995). A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection. *Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence*, 1137–1143. San Francisco, CA: Morgan Kaufmann.

Levin, B. (1993). *English Verb Classes and Alternation: A Preliminary Investigation*. Chicago: The University of Chicago Press.

Lewis, D. D. (1998). Naive Bayes at forty: The independence assumption in information retrieval. In Nédellec, C. & Rouveirol, C. (Eds.), *Proceedings of Tenth European Conference on Machine Learning*, 1398, 4–15, Heidelberg: Springer-Verlag.

***Longman Dictionary of Contemporary English*. (1995).** Third Edition. Essex, England: Longman Group Ltd.

Mel'čuk, I. A. (1974). *A Theory of the Meaning-Text Type Linguistic Models*. (In Russian). Moscow: Nauka Publishers.

Mel'čuk, I. A. (1995). Phrasemes in Language and Phraseology in Linguistics. In Everaert, M., van der Linden, E.-J., Schenk, A. & Schreuder, R. (Eds.), *Idioms: Structural and Psychological Perspectives*, 167–232. Hillsdale, NJ: Lawrence Erlbaum.

Mel'čuk, I. A. (1996). Lexical Functions: A Tool for the Description of Lexical Relations in a Lexicon. In Wanner, L. (Ed.), *Lexical Functions in Lexicography and Natural Language Processing*, 37–102. Amsterdam, Philadelphia, PA: Benjamins Academic Publishers.

Merriam-Webster Open Dictionary. Available at: http://www3.merriam-webster.com/opendictionary/

Mitchell, T. (1997). *Machine Learning*. McGraw Hill.

Pedersen, T. (2000). A simple approach to building ensembles of naive bayesian classifiers for word sense disambiguation. *Proceedings of the First Annual Meeting of the North American Chapter of the Association for Computational Linguistics*, 63–69. Seattle, WA.

Pranckeviciene, E., Somorjai, R. & Tran, M.N. (2007). Feature/model selection by the linear programming svm combined with state-of-art classifiers: What can we learn about the data. *Proceedings of the 20th International Joint Conference on Neural Networks*, 1422–1428.

Real Academia Española. (2001). *Diccionario de la Lengua Española*. Madrid: Real Academia Española.

Sidorov, G. (1996). Lemmatization in automatized system for compilation of personal style dictionaries of literature writers. In: *Word of Dostoyevsky*, 266–300. Moscow, Russia: Russian Academy of Sciences.

***Spanish Web Corpus*.** Available at http://trac.sketchengine.co.uk/wiki/Corpora/SpanishWebCorpus/

***Spanish WordNet*.** Available at http://www.lsi.upc.edu/~nlp/web/index.php?Itemid=57&id=31&option=com_content&task=view

The University of Waikato Computer Science Department Machine Learning Group. WEKA download at http://www.cs.waikato.ac.nz/~ml/weka/index_downloading.html

Vossen P. (Ed.). (1998). *EuroWordNet: A Multilingual Database with Lexical Semantic Networks*. Dordrecht: Kluwer Academic Publishers,

Wanner, L. (2004). Towards automatic fine-grained classification of verb-noun collocations. *Natural Language Engineering*, 10(2), 95–143. Cambridge: Cambridge University Press.

Wanner, L., Bohnet, B. & Giereth, M. (2006). What is beyond Collocations? Insights from Machine Learning Experiments. *Proceedings of EURALEX*.

Witten, I. H. & Frank, E. (2005). *Data Mining: Practical machine learning tools and techniques*. Second Edition. San Francisco: Morgan Kaufmann

Biographic Information

Prof. **Alexander Gelbukh** holds M.Sc. degree in mathematics y Ph.D. degree in computer science. Since 1997 he leads the Natural Language Processing of the Computing Research Center of the National Polytechnic Institute (CIC-IPN), Mexico. He is academician of the Mexican Academy of Sciences, National Researcher of Mexico of excellence level 2, and the Secretary of the Mexican Society for Artificial Intelligence (SMIA). He is author or editor of more than 440 publications and co-author of three books in the areas of natural language processing and artificial intelligence. More information about him can be found on his personal page www.Gelbukh.com.

Olga Kolesnikova obtained her M.Sc. degree in Linguistics from Novosibirsk State Pedagogical Institute, Russia, in 1989. At present she is a Ph.D. student at the Center for Computing Research of National Polytechnic Institute (CIC-IPN), Mexico. Her research is in the area of computational linguistics; in particular, she is interested in text semantic analysis.