Procesamiento del Lenguaje Natural, Revista nº 69, septiembre de 2022, pp. 141-152 recibido 31-03-2022 revisado 11-05-2022 aceptado 19-05-2022

Semantic Relations Predict the Bracketing of Three-Component Multiword Terms

Las Relaciones Semánticas Predicen la Desambiguación Estructural de las Unidades Terminológicas Poliléxicas con Tres Formantes

Juan Rojas-Garcia
University of Granada, Granada, Spain
juanrojas@ugr.es

Abstract: For English multiword terms (MWTs) of three or more constituents (e.g., *sea level rise*), a semantic analysis, based on linguistic and domain knowledge, is necessary to resolve the dependency between components. This structural disambiguation, often known as bracketing, involves the grouping of the dependent components so that the MWT is reduced to its basic form of modifier+head, as in [*sea level*] [*rise*]. Knowledge of these dependencies facilitates the comprehension of an MWT and its accurate translation into other languages. Moreover, the resolution of MWT bracketing provides a higher overall accuracy in machine translation systems and sentence parsers. This paper thus presents a pilot study that explored whether the bracketing of a ternary compound, when used as an argument in a sentence, can be predicted from the semantic information encoded in that sentence. It is shown that, with a random forest model, the semantic relation of the MWT to another argument in the same sentence, the lexical domain of the predicate, and the semantic role of the MWT were able to predict the bracketing of the 190 ternary compounds used as arguments in a sample of 188 semantically annotated sentences from a Coastal Engineering corpus (100% F_1-score). Furthermore, only the semantic relation of an MWT to another argument in the same sentence proved enormous capability to predict ternary compound bracketing with a binary decision-tree model (94.12% F_1-score).

Keywords: Semantic Relation, Multiword-Term Bracketing, Random Forest, Decision Tree.

Resumen: En unidades terminológicas poliléxicas (UTP) con tres o más formantes en lengua inglesa (p.ej., *sea level rise*), establecer la dependencia entre dichos formantes requiere de un análisis lingüístico y de conocimiento especializado del área concreta en que se emplean las UTP. Esta desambiguación estructural, o *bracketing*, implica el agrupamiento de los formantes para reducir la UTP a su estructura básica de modificador+núcleo, como en [*sea level*] [*rise*]. Conocer el *bracketing* de una UTP no solo facilita su comprensión y traducción a otras lenguas, sino que también mejora el desempeño de los sistemas de traducción automática y de los analizadores sintácticos. Por tanto, en este artículo presentamos un estudio piloto que explora si el *bracketing* de una UTP con tres formantes, al emplearse como argumento en una oración, puede predecirse a partir de la información semántica codificada en dicha oración. Se muestra que, con un modelo random forest, la relación semántica de la UTP con otro argumento en la misma oración, el dominio léxico del verbo y el rol semántico de la UTP son capaces de predecir el *bracketing* de las 190 UTP ternarias que se usan como argumento en una muestra de 188 oraciones, anotadas semánticamente y extraídas de un corpus sobre ingeniería de costas (con un valor de F_1 del 100 %). Además, únicamente la relación semántica que mantiene una UTP ternaria con otro argumento en la misma oración posee una enorme capacidad para predecir su *bracketing* mediante un árbol de decisión binario (con un valor de F_1 del 94,12 %).

Palabras clave: Relación Semántica, Desambiguación Estructural de Unidades Terminológicas Poliléxicas, Random Forest, Árbol de Decisión.

ISSN 1135-5948. DOI 10.26342/2022-69-12

1 Introduction

A set of 1,694 sentences from a Coastal Engineering corpus, in which a named river (e.g., Salinas River) was an argument of the predicate of the sentences, were semantically analyzed and annotated with the semantic relation between the arguments, the lexical domain of the predicates, and the semantic role of the arguments.

This paper presents the statistical analysis of those semantic annotations with a view to finding evidence that the structural disambiguation, or bracketing, of a three-component multiword term (e.g., [*sand supply*] [*decrease*]) can be predicted from the semantic information encoded in the sentence where the ternary compound is used as an argument. For this experiment, we assumed that the context, which constrains the factors that drive understanding (Leech G., 1981), also helps to resolve the structural disambiguation of a ternary compound. This assumption comes from the daily experience of a translator who must deal with ternary compounds in a specialized text. Although the compounds are somewhat familiar, it is useful to craft definitions for them to facilitate their translation into another language based on their context of use.

The rest of this paper is organized as follows. Section 2 presents a fundamental background of bracketing of multiword terms. Section 3 provides a literature review of predictive models for bracketing, mostly from the perspective of variables and resources used for the task of compound bracketing prediction. Section 4 explains the materials used in this study. Section 5 covers our semantic approach to predicting ternary compound bracketing based on two supervised models, namely decision tree, and random forest. Also described are the sample of ternary compounds, the training and testing phases for the predictive models, and the results, which provide linguistic insights as to how semantic relations, predicate lexical domains, and semantic roles are intertwined with the bracketing of ternary compounds. Section 6 discusses the results and compares them to those outlined in the literature review. Finally, Section 7 presents the conclusions derived from this work along with plans for future research.

2 Bracketing of Multiword Terms

When multiword expressions are used in specialized domains, they are known as multiword terms (MWTs). MWTs often have more than two components. For instance, in Coastal Engineering, *beach size sand supply* refers to the supply of sand, usually provided by rivers, whose grain size is appropriate to mitigate beach erosion. The most frequent MWTs in specialized texts are endocentric because they specify a broader concept or hypernym. For example, *beach size sand supply* is a type of *sand supply* since the grain size of the sand is specified. It is thus the dimension activated to form the hyponym.

For MWTs of three or more constituents, a semantic analysis, based on linguistic and domain knowledge, is necessary to resolve the dependency between components. This structural disambiguation, often known as *bracketing* or *parsing*, involves the grouping of the dependent components so that the MWT is reduced to its basic form of modifier+head, as in [*beach size*] [*sand supply*]. Knowledge of these dependencies facilitates the comprehension of an MWT and, consequently, its accurate translation into other languages.

Therefore, before including MWTs in terminological knowledge bases, it is often necessary to structurally disambiguate them to make their relational structure explicit and thus favor knowledge acquisition (León-Araúz P. et al., 2021). Furthermore, the resolution of MWT bracketing provides a higher overall accuracy in machine translation systems (Green N., 2011), sentence parsers (Vadas D. and Curran J.R., 2008), and in systems aimed at determining the implicit semantic relation holding between modifier and head in MWTs of three or more components (Kim S.N. and Baldwin T., 2013).

3 Review of Bracketing Prediction Methods

Previous work on compound parsing/bracketing exploits either unsupervised methods (e.g., based on bigram corpus frequency) or supervised ones (i.e., based on training data, containing manually parsed/bracketed compounds, which are used to train an algorithm for predicting compound bracketing).

The two basic unsupervised approaches are the adjacency model (Marcus M., 1980; Pustejovsky J. et al., 1993), and the dependency model (Lauer M., 1994). For a ternary compound such as *sea level rise* (i.e., increase in sea level), the adjacency model concludes whether *level* is more closely associated with *sea* (leading to a left-branched structure) or to *rise* (leading to a right-branched structure). In contrast, the dependency model resolves whether *sea* is more strongly associated with *level* (leading to a left-branched structure) or with *rise* (leading to a right-branched structure). In this case, the correct bracketing of *sea level rise* is left-branched. The way of measuring the association strength between two of the words (or constituents) in the compound is based on association measures estimated from corpus data, such as bigram frequency, point-wise mutual information, or chi-squared, among others.

Resnik P.S.'s (1993) method for ternary compounds, based on the adjacency model and the association measure called *selectional association*, estimated from the parsed *Wall Street Journal* corpus (30 million words), achieved an overall accuracy of 72.6% (with a sample of 157 ternary compounds from the Penn Treebank corpus, 64.1% left-branched, and 35.9% right-branched). In contrast, Lauer M. (1995) adopted the dependency model for his method, based on the ratio of left- to right-bracketing probability for a ternary compound, estimated from *Grolier's Encyclopedia* (8 million words). The author calculated probabilities of conceptual categories in the taxonomy underlying *Roget's Thesaurus* (Roget P.M., 1852)[1], rather than for individual words, to avoid data sparsity problems. His method reached an overall accuracy of 80.7% (with a sample of 244 ternary compounds from *Grolier's Encyclopedia*, 66.8% left-branched, and 33.2% right-branched).

Nakov P. and Hearst M. (2005) developed an unsupervised, knowledge-rich method for parsing ternary compounds. Their approach included:

(1) Ten types of surface variable, such as dashes (e.g., *beach-sand transport* points to a left-bracketed compound), possessive markers (e.g., *city's water supply* indicates a right-bracketed compound), and acronyms (e.g., *pH quality control (QC)* reveals a right-bracketed compound).

(2) Three types of paraphrase variable, namely prepositional phrases (e.g., *distance <u>from the</u> river mouth* means that *river mouth distance* is left-bracketed), copula paraphrases (e.g., *water product <u>that/which is a</u> mixture* proves that *mixture water product* is right-bracketed), and verbal paraphrases (e.g., *impacts <u>associated with</u> river pollution* implies that *river pollution impact* is left-bracketed).

The authors concluded that the adjacency and dependency models showed comparable performance when using the chi-squared association measure and the number of web search engine page hits for approximating corpus frequencies, as suggested by Lapata M. and Keller F. (2004). Although their method achieved an overall accuracy of 95.35%, this result was probably biased toward the majority left-bracketing class because of the bracketing-imbalanced sample of 430 ternary compounds from a corpus of biomedical domain abstracts retrieved from MEDLINE (84% left-branched, and only 16% right-branched).

Girju R. et al. (2005) implemented a supervised model for bracketing ternary compounds with the machine-learning technique decision tree. They employed a total of 15 semantic variables based on WordNet senses, five variables for each compound constituent, namely the top three WordNet semantic categories for each constituent, derivationally-related forms, and whether the constituent was a nominalization. The algorithm reached an overall accuracy of 83.10%, with a sample of 728 ternary compounds from the *Wall Street Journal* component of the Penn Treebank corpus (Marcus M. et al., 1993), 67.4% left-branched, and 32.6% right-branched.

Kim S.N. and Baldwin T. (2013) devised a method that consisted of automatically determining the semantic relations between the pairs of words in a ternary compound, and then predicting bracketing from the constituent pair whose semantic relation coincided with that of the ternary compound. When this method was combined with that of Nakov P. and Hearst M. (2005), it achieved an overall accuracy of 74.1% with a sample of 1,571 ternary compounds from the *Wall Street Journal* corpus. However, no information was provided regarding the percentage of left- and right-bracketing within the sample.

The supervised method by Bergsma S. et al. (2010) used both n-gram variables (the logarithm of the frequency of all constituent subsets appearing in the Google V2 corpus), and Boolean lexical variables that indicated the presence or absence of a particular string at a given position in the compound (the constituents and their position, the entire ternary compound, as well as a capitalization pattern of the constituent sequence). As a machine-learning technique, the authors applied the support-vector machine algorithm, which reached an overall accuracy of 91.6%, with a sample of 2,150 ternary compounds from the *Wall Street Journal* corpus (70.5% left-branched, and 29.5% right-branched).

Vadas D. and Curran J.R. (2007) developed a supervised method for parsing ternary compounds based on the machine-learning technique of logistic regression. They used 88,568 variables, an extremely large number, which can be summarized as follows:

(1) Bigram frequencies were collected from two sources, namely hit counts from the web search engine Google, and frequencies in the Google Web 1T corpus (Brants T. and Franz A., 1993).

(2) The pairs of compound constituents, and the surface variables by Nakov P. and Hearst M. (2005), were compared according to both the adjacency and dependency models by means of the chi-squared, and bigram probability association measures.

[1] http://www.gutenberg.org/ebooks/10681.

(3) Lexical features for all unigrams and bigrams in a ternary compound, along with their position within the compound.

(4) Contextual variables, consisting of bag-of-word features for both the words in the sentence where the compound is used, and for a two-word window on each side of the compound.

(5) For every n-gram and context window feature, their part-of-speech tags and named entity tags were added.

(6) For each sense of each constituent in the ternary compound, a semantic feature for its synset, as well as the synset of each of its hypernyms up to the root, were extracted from WordNet, and incorporated into the supervised model as additional variables.

This method achieved an F_1-score of 93.01%, with a sample of 5,582 ternary compounds from the Penn Treebank corpus (58.99% left-branched, and 41.01% right-branched).

The supervised system by Pitler E. et al. (2010) was able to bracket compounds of three or more constituents (including the conjunction *and*). Applying the support-vector machine algorithm, the system first calculated the probability that a word sequence, within a compound, was a constituent, given the entire compound as context. Then, using these probabilities, the system predicted the bracketing of a compound with the CYK parser (i.e., Cocke-Younger-Kasami algorithm). As variables for the system, the authors employed:

(1) The position of the proposed bracketing within the compound.

(2) The association measure point-wise mutual information (PMI) between all word pairs in the compound, derived from the Google V2 corpus (Lin D. et al., 2010).

(3) Boolean lexical variables to indicate the presence of a particular word at each position in the compound.

(4) Boolean variables to inform about the shape of the compound, namely the presence of capitalized letters and hyphenated words provided information concerning the possibility that the compound included a named entity.

The system reached an overall accuracy of 95.4%, with a sample of 64,844 compounds of three or more constituents from the Penn Treebank corpus, but bracketing-related information in the form of percentages was not provided.

Lazaridou A. et al. (2013) tackled the parsing of a ternary compound using a *semantic plausibility measure* derived from a distributional semantic model trained on a corpus of 2.8 billion tokens, where the vector of a ternary compound was obtained from the combination of the vectors of each of its constituents. This supervised method relied on the support-vector machine algorithm with 14 variables, summarized as follows: (1) 12 variables for representing the semantic plausibility of either the left- or right-bracketing; (2) two variables for the PMI values of the word pairs in the compound, according to the adjacency model. The method achieved an overall accuracy of 85.6%, with a sample of 2,227 ternary compounds from the Penn Treebank corpus (34.4% left-branched, and 65.6% right-branched).

Faruqui M. and Dyer C. (2015) also addressed the ternary compound bracketing with word vectors. However, their semantic model was non-distributional because the vectors did not encode any word co-occurrence information. Instead, the vector dimensions were Boolean variables that represented linguistic knowledge derived from resources such as WordNet (Fellbaum C.A., 1998), FrameNet (Ruppenhofer J. et al., 2010), and Penn Treebank. As such, the vector length for a single word included a total of 172,418 dimensions. The vector of a ternary compound was then obtained by appending the vector of each constituent, which resulted in a ternary compound vector of 517,254 dimensions. This combined vector was the input of the machine-learning technique of logistic regression, which achieved an overall accuracy of 83.3% in the same sample of ternary compounds collected by Lazaridou A. et al. (2013).

For the unsupervised method by Ménard P.A. and Barrière C. (2014), the usage of different resources for the bracketing of compounds of three and more constituents was compared, namely the English Google Web N-grams (Lin D. et al., 2010), English Google Books Ngrams (Michel J.B. et al., 2010), and open linked data DBpedia (Hellmann S. et al., 2009). The association measures chi-squared, PMI, and Dice, and the number of valid DBPedia paths were also analyzed. Their algorithm created an initial list containing all of the word pairs from a compound, which were then sorted in descending order of association scores. A second list of dependencies, which defined the complete bracketing of the compound, was constructed from the first list. For ternary compounds, the method with the English Google Books N-grams and the PMI achieved the highest overall accuracy, with a value of 81.47% on a sample of 2,889 ternary compounds from the Penn Treebank corpus (79.2% left-branched, and 20.8% right-branched).

Similarly, for the bracketing of compounds of three and more constituents, Barrière C. and Ménard P.A. (2014) applied the unsupervised method of Ménard P.A. and Barrière C. (2014), but relied on a word association model that combined the lexical, relational, and coordinate nature of the associations between all pairs of words within a compound. The information for their word association model was collected from Wikipedia. The system reached an overall accuracy of 73.16%, with a sample of 4,749 compounds of three and more constituents from the Penn Treebank corpus, but the specific accuracy for the subset of ternary compounds was not provided.

León-Araúz P. et al. (2021) developed an unsupervised, knowledge-rich method for bracketing specialized ternary compounds in the domain of wind energy. The authors used 12 variables, mainly related to the surface and paraphrase variables proposed by Nakov P. and Hearst M. (2005), which measured frequency counts in a specialized corpus on wind energy. The counts were collected by means of CQL (Corpus Query Language) queries in the Sketch Engine corpus manager. A total of 34 specific CQL queries were designed for the extraction of occurrences of each of the linguistic structures underlying the 12 variables. Based on the results, the authors formulated 16 rules to decide on the bracketing of a ternary compound. Hence, the final bracketing structure was decided by applying the majority vote strategy to the votes of the individual rules. As such, the CQL queries and rules permitted the implementation of a system to automate the compound bracketing task for users such as translators and terminologists. The method achieved an overall accuracy of 86.4%, with a sample of 103 ternary compounds from the wind energy domain (67% left-branched, and 33% right-branched).

In short, previous research focused on semantic information provided by the components of an MWT. The number of variables used for prediction ranged from 12 to 517,254 features. These variables were mostly based on n-gram statistics, and semantic information of the MWT components stored in linguistic resources such as WordNet. The overall accuracy of the prediction models ranged from 72.60% to 95.40%.

Our approach, however, was based on semantic information that previous research has not as yet considered. This semantic information was encoded in both the co-text of a ternary compound (i.e., the sentence where the ternary compound was used as an argument) and the ternary compound seen as a unit (i.e., its semantic role). The set of predictor variables consisted of only three (i.e., the semantic relation, predicate lexical domain, and semantic role of the MWT), whereas previous research employed a minimum of 12 variables (León-Araúz P. et al., 2021).

4 Materials

A set of 1,694 sentences, in which a named river (e.g., Mississippi River) was an argument of the predicate of the sentences, were semantically analyzed and annotated. These sentences were extracted from a subcorpus of English texts on Coastal Engineering, comprising roughly 7 million tokens and composed of specialized texts (scientific articles, technical reports, and PhD dissertations), and semi-specialized texts (textbooks and encyclopedias on Coastal Engineering). This subcorpus is part of the English EcoLexicon Corpus (23.1 million words) (see León-Araúz P. et al. (2018) for a detailed description).

5 Semantic Approach for MWT Bracketing

Since the semantic information in a sentence firmly guides its syntactic parsing (Fillmore C.J., 1968; Lazaridou A. et al., 2013), one could assume that the correct bracketing of an MWT, when used as an argument in a sentence, can be predicted from the semantic information encoded in that sentence. In other words, the context, which constrains the factors that drive understanding (Leech G., 1981), helps to resolve the structural disambiguation of the ternary compound.

As semantic information in a sentence, this pilot study explored the contribution of three semantic variables to the prediction of ternary compound bracketing. These variables were the lexical domain of the verb, semantic role of the ternary compound, and semantic relation of the ternary compound to the named river. From the 1,694 sentences semantically analyzed and annotated, 188 sentences contained 190 ternary compounds as arguments. This sample of 190 ternary compounds, along with the values of the abovementioned three semantic variables annotated in their corresponding sentences, were employed for the training and testing of two supervised models to predict whether a ternary compound was right-branched or left-branched.

5.1 Annotation of the Semantic Variables

A set of 1,694 sentences from the corpus, where 294 different rivers are mentioned, were annotated by three terminologists from the LexiCon research group of the University of Granada (Spain). They performed the semantic annotation of the predicate-argument structure of a sentence by assigning a: (1) lexical domain to the predicate; (2) semantic role to the arguments of the predicate; (3) semantic relation to the link between the named river and the other arguments in the sentence; and (4) bracketing (left or right) to the ternary compounds used as arguments in the sentence. The values of these four semantic variables are shown in Table 1.

Semantic variables annotated	Values
Lexical domain of the predicates (8 values)	CHANGE, MOVEMENT, EXISTENCE, POSSESSION, POSITION, MANIPULATION, ACTION, COGNITION
Semantic roles of the arguments (13 values)	AGENT, RESULT, PATIENT, THEME, LOCATION, RECIPIENT, INSTRUMENT, TIME, RATE, MANNER, DESCRIPTION, CONDITION, PURPOSE
Semantic relation between the ternary compound and the named river (30 values)	*type_of*, *part_of*, *made_of*, *delimited_by*, *located_at*, *takes_place_in*, *phase_of*, *affects*, *causes*, *result_of*, *attribute_of*, *has_function*, *studies*, *measures*, *effected_by*, *improves*, *worsens*, *creates*, *becomes*, *gives*, *gives_to*, *receives*, *receives_from*, *drains*, *has_path*, *transfers*, *discharges_into*, *places*, *controls*, *applied_to*
Bracketing of the ternary compounds in the sentences (2 values)	RIGHT, LEFT

Table 1: Semantic variables annotated in the set of sentences, and their values.

The most frequent verbs in the corpus are general language verbs (e.g., *accumulate*, *pollute*, *increase*, *discharge*, *supply*, *drain*), which are also used in specialized texts and thus reflect how environmental entities interact. In this sense, such verbs are susceptible to classification in the lexical domains proposed by Faber P. and Mairal R. (1999), within the Functional Lexematic Model. These lexical domains were used to annotate the predicates of our set of sentences, and shown in Table 1.

Specialized knowledge representation includes semantic properties that help to describe the nature of entities and processes. These semantic properties are reflected as the relations between a predicate and its arguments, which are typical semantic roles. The semantic roles used to annotate the arguments in our set of sentences largely coincided with those specified by Kroeger P.R. (2005: 54-55), and Thompson P. et al. (2009), and summarized in Table 1.

Conceptual description of specialized concepts includes their relational behavior. These relations, depicted by Faber P. et al. (2009) for environmental concepts, with additional non-hierarchical relations specific to named rivers (Rojas-Garcia J., forthcoming), were all used to annotate the semantic relation between the arguments in our set of sentences, and collected in Table 1.

The inter-annotation agreement coefficient, *Cohen's kappa* (κ), showed a very good agreement for all the annotator pairs (κ>90%, *p*-value<0.05) in the annotation of the semantic roles, relations, and bracketing according to Krippendorff K.'s (2012) recommendations for text content analysis. Notwithstanding, the disagreements in the original annotations were resolved based on discussion between the annotators to reach a consensus on the definitive annotations of semantic roles, relations, and bracketing.

For the initial annotation of predicates with lexical domains, the inter-annotation agreement was lower for all the annotator pairs (84%<κ<88%, *p*-value<0.05), indicating that this variable lent itself to alternative, though plausible, interpretations. A review of the differences between annotators showed that the lexical domains of MOVEMENT and POSSESSION were more prone to confusion. The issues fundamentally arose from verbs that could potentially belong to more than one lexical domain (e.g., *drain* and *discharge*), as Faber P. and Mairal R. (1999) already proved. To arrive at a consensus on the definitive annotations of lexical domains, the factorization of meaning from the Functional Lexematic Model framework was applied to verbs to resolve disagreements between the annotators.

5.2 Description of the Sample of MWTs

A selection of 10 sentences from the sample, which incorporated ternary compounds as arguments, is provided in Table 2. For each of those 10 sentences, Table 3 shows the values of the following four annotated variables: (1) lexical domain of the predicate (*LexDom*); (2) semantic role of the ternary compound (*SemRol_mwt*); (3) semantic relation between the ternary compound and the named river (*SemRel*); and (4) bracketing of the ternary compound (*Bracketing*), which was the variable to be predicted.[2]

The distribution of bracketing structures within the MWT sample was reasonably balanced between left-branching (110 MWTs, 58% of the sample), and right-branching (80 MWTs, 42% of the sample). Table 4 summarizes the counts for the sample data, disaggregated by lexical domain and bracketing structure of the MWTs, and describes the distribution of the 190 MWTs across these variables. Some conclusions could be drawn from the characteristics of the sample: (1) sentences whose predicate belonged to the lexical domains of MOVEMENT, ACTION, POSITION, MANIPULATION, and COGNITION included ternary compounds which were only right-branched; and (2) sentences whose predicate belonged to the lexical domain of POSSESSION incorporated ternary compounds which were only left-branched.

[2] The whole dataset of MWTs, the values of the annotated variables, and the corpus will be available on the website of the LexiCon research group of the University of Granada (Granada, Spain) (http://lexicon.ugr.es/).

Sentences from the Sample with Ternary Compounds as Arguments
(1) Blackstone River draining into Narragansett Bay has been extensively dammed, and although not well quantified, models show **decreasing sediment load** in the **Blackstone River**.
(2) The dramatical sediment load variation in the **Pearl River**, with the almost unchanged **water discharge level**, represents an example of such effect that human activities can have on river deltas.
(3) **Muddy silt deposition** in the **Clyne River** discharging into the Swansea Bay would increase.
(4) **Rising sea levels** change **Salinas River Estuary** and could thus potentially alter sediment supplies and process patterns.
(5) The **Salinas River** no longer contributes substantial **beach size sand** to the Littoral Cell because the river gradient has greatly decreased with sea level rise, reducing the flow rate.
(6) The **River Murray** flows across Tertiary formations to enter coastal lagoons behind the **dune calcarenite barriers** of Encounter Bay.
(7) Not all the sediments drained by the **Dee River** participate to **coastal sediment transport**.
(8) The field site for this study is the **Zuidgors salt marsh**, located in the **Western Scheldt estuary** in The Netherlands.
(9) **Natural sediment supply** within this region is defined by the **Ventura River** that drains large watersheds.
(10) The **average discharge rate** of beach size sand in the **Salinas River** is estimated at approximately 65,000 cubic yards per year.

Table 2: Selection of 10 sentences (from the sample of 188 sentences), which included 10 ternary compounds as arguments.

MWT	**LexDom**	**SemRol_mwt**	**SemRel**	**Bracketing**
decreasing [sediment load]	EXISTENCE	DESCRIPTION	*attribute_of*	RIGHT
[water discharge] level	EXISTENCE	THEME	*attribute_of*	LEFT
[muddy silt] deposition	CHANGE	PATIENT	*takes_place_in*	LEFT
rising [sea level]	CHANGE	AGENT	*worsens*	RIGHT
[beach size] sand	POSSESSION	THEME	*gives*	LEFT
dune [calcarenite barrier]	MOVEMENT	AGENT	*has_path*	RIGHT
coastal [sediment transport]	ACTION	DESCRIPTION	*affects*	RIGHT
Zuidgors [salt marsh]	POSITION	THEME	*located_at*	RIGHT
natural [sediment supply]	MANIPULATION	PATIENT	*controls*	RIGHT
average [discharge rate]	COGNITION	THEME	*attribute_of*	RIGHT

Table 3: Semantic annotations and variables for a set of 10 MWTs out of the 190 MWTs that comprised the sample. The semantic information in the rows corresponds to the respective sentences in Table 2.

Lexical Domain	**LEFT-branched MWTs**	**RIGHT-branched MWTs**	**Total**
MOVEMENT	0	10	10 (5.3%)
POSSESSION	30	0	30 (15.8%)
CHANGE	20	10	30 (15.8%)
EXISTENCE	60	20	80 (42.0%)
ACTION	0	10	10 (5.3%)
POSITION	0	10	10 (5.3%)
MANIPULAT.	0	10	10 (5.3%)
COGNITION	0	10	10 (5.3%)
Total	**110 (58%)**	**80 (42%)**	**190 (100%)**

Table 4: Description of the sample of ternary compounds.

5.3 Supervised Models

Regarding the supervised models for classification, *binary decision tree* and *random forest* were tested to predict ternary compound bracketing. Since variables in our dataset were categorical, both tree-based models were adopted because they can efficiently manage qualitative variables (James G. et al., 2015: 315).

A decision-tree model is simple and readily interpretable because the set of prediction rules is graphically summarized in a tree, typically drawn upside down, in the sense that the terminal nodes or leaves, which convey the predictions, are at the bottom of the tree. However, it is usually not competitive with other predictive models.

For that reason, we also experimented with a random forest model, which produces a large number of decision trees, and then combines them to reach a single consensus prediction. Namely, each tree in the ensemble (or forest) casts a vote for the bracketing of an MWT, which is finally classified into the bracketing structure that has the most votes. Random forest models thus lead to remarkable improvements in prediction accuracy, at the expense of loss in interpretation since it is difficult to obtain insight as to how the model makes the predictions.

5.4 Data Splitting

For the construction and evaluation of the models, the dataset with the 190 MWTs was divided into two: (1) the training dataset to create the models (with 133 MWTs, 70% of the original dataset), and (2) the test dataset to qualify model performance (57 MWTs, 30% of the original dataset).

For both the training and test datasets to have the same distribution in the outcome variable (i.e., *Bracketing*) as the original dataset (i.e., 58% left-branched MWTs, and 42% right-branched MWTs), stratified random sampling was conducted, which randomly sampled observations within the classes LEFT and RIGHT of the *Bracketing* variable in the original dataset.

5.5 Model Performance Measures

The quality of the two models (decision tree and random forest) was assessed by analyzing how well they performed on the test dataset, which was hidden from the model-building process for evaluation purposes. As such, the predictions of the models were compared to the true classes of the test dataset (i.e., the true bracketing structures LEFT and RIGHT, recorded in the *Bracketing* variable of the test dataset), and performance measures were calculated.

A widely used performance measure is *overall accuracy*, which provides the percentage of correctly classified instances. However, this measure has some drawbacks in imbalanced datasets, or datasets whose outcome variable exhibits a significant disproportion among the number of instances of each class.

According to Fernández A. et al. (2018: vii), the learning process of most classification algorithms, including decision tree and random forest, is often biased toward the majority-class instances, and minority-class ones are thus not well modelled into the final system. Consequently, in imbalanced scenarios, the accuracy measure may mask a poor classification performance in the minority class. Unfortunately, as already seen in the literature review, there is much research on bracketing prediction that still uses overall accuracy with severely bracketing-imbalanced datasets. Therefore, despite the fact that our dataset was only slightly bracketing-imbalanced, we preferred to use, in addition to accuracy, other measures that were not sensitive to disparities in the class proportions to evaluate classification performance. Such measures were the *area under the ROC curve*, and the *F_1-score* (Fernández A. et al., 2018: 52-55).

The *receiver operating characteristic (ROC) curve* is a function of the sensitivity and specificity of a two-class predictive model to evaluate its trade-off between both measures. *Sensitivity* is the fraction of the minority-class instances (in our case, the right-branched MWTs) that are correctly classified, whereas *specificity* refers to the proportion of the majority-class instances (in our case, the left-branched MWTs) that are correctly classified. Hence, the *a*rea *u*nder the ROC *c*urve (henceforth referred to as AUC) is a method for combining sensitivity and specificity into a single value. AUC ranges from 0 to 1. The higher the AUC, the better the performance of the model at distinguishing between the two classes.

The F_1-score is the harmonic mean between the precision and recall of a predictive model. *Precision* is the fraction of correctly classified minority-class instances among the instances classified as belonging to the minority class, whereas *recall* is the same as sensitivity. Thus, the F_1-score evaluates the trade-off between correctness and coverage in classifying minority-class instances.

5.6 Construction of the Predictive Models

The predictors *SemRel*, *LexDom*, and *SemRol_mwt* were used to construct two predictive models with the *caret* package (Kuhn M., 2021) for the R programming language.

For the random forest, 7-fold cross-validation in the training dataset was used to evaluate its performance in training. Although 10 folds are conventionally employed, we chose 7 folds, a divisor of 133, so that the number of instances in all folds would be the same (i.e., 19 instances). During the process of tuning parameters, the AUC performance measure was chosen to be maximized. Accordingly, the random forest model attained in training an AUC value equal to 1.0 when: (1) the splits in the trees were allowed to use one predictor of a subset of one predictor; and (2) the number of trees in the forest was, surprisingly, only three trees. In the test dataset, the random forest also achieved an AUC value equal to 1.0. Consequently, the three predictors were capable of correctly predicting bracketing in the test dataset with a random forest model.

Similarly, for the decision tree, 7-fold cross-validation in the training dataset was employed to evaluate its performance in training. During the process of tuning parameters, the AUC performance measure was also chosen to be maximized. Therefore, the decision-tree model yielded in training the greatest AUC, equal to 0.9545, when: (1) the cost-complexity parameter (*cp*) was equal to *cp*=0.8392857; and (2) the splitting criterion for predictors was the *information gain*, and not the *Gini index*. In the test dataset, the decision-tree model achieved an AUC value also equal to 0.9545, which indicated a very satisfactory performance.

Table 5 provides further performance measures, in the training and test datasets, for the random forest and decision-tree models.

Predictors: *SemRel*, *LexDom*, and *SemRol_mwt*					
	Decision Tree Model				
Dataset	**AUC**	**Precision**	**Recall**	**F_1**	**Accura.**
Train	0.9545	0.8952	1.000	0.9430	0.9474
Test	**0.9545**	0.8889	1.000	**0.9412**	**0.9474**
	Random Forest Model (3 ensembled decision trees)				
Train	1.0000	1.0000	1.0000	1.0000	1.0000
Test	**1.0000**	1.0000	1.0000	**1.0000**	**1.0000**

Table 5: Performance measures of the models for bracketing prediction with the predictors semantic relation, lexical domain, and semantic role.

Since the decision-tree model reached a significant AUC in the test dataset (AUC=0.9545), its only prediction rule, graphically summarized in Figure 1, is worth mentioning.

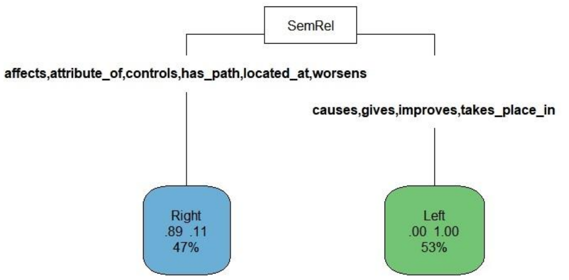
SemRel
affects,attribute_of,controls,has_path,located_at,worsens
causes,gives,improves,takes_place_in
Right
.89 .11
47%
Left
.00 1.00
53%

Figure 1: Classification tree for bracketing prediction, inferred by the decision-tree model trained with the predictors semantic relation, lexical domain, and semantic role of the MWTs.

In our constrained context (i.e., specialized ternary compounds from Coastal Engineering, used in sentences where a named river was mentioned), the classification tree of the model, displayed in Figure 1, can be interpreted as follows.

SemRel was the most important factor in determining *Bracketing*, and the only predictor selected by the decision-tree model. In our opinion, the predictive power of the semantic relation between an MWT and another argument in the same sentence is so high that the model was obliged to reject the use of the predictors *LexDom* and *SemRol_mwt* to avoid overfitting to training data.

As such, the ternary compounds whose semantic relation to the other argument, filled with a named river in our case, belonged to the group formed by *causes*, *gives*, *improves*, and *takes_place_in* (right-hand branch in the classification tree) accounted for 53% of the sample; these MWTs were all left-branched and correctly classified. It thus seemed that these four semantic relations forced the use of only left-branched MWTs.

In contrast, the ternary compounds whose semantic relation to the other argument fell into the group formed by *affects*, *attribute_of*, *controls*, *has_path*, *located_at*, and *worsens* (left-hand branch in the classification tree) comprised 47% of the sample, and could be right- or left-branched; under these conditions, the model correctly classified all the right-branched MWTs (89%), but misclassified the true left-branched MWTs (11%) as right-branched.

An analysis of the errors made by the decision-tree model revealed that, both in the training and test datasets, those left-branched MWTs with the values *SemRel*=*attribute_of*, *LexDom*=EXISTENCE, and *SemRol_mwt*=THEME (e.g., *water discharge level*, in row 2 of Table 3), were all misclassified as right-branched.

5.7 Baseline Models

The results of our semantic approach were compared to those of four baseline models, namely: (1) adjacency model with the point-wise mutual information (PMI) association measure, as defined by Marcus M. (1980); (2) adjacency model with the chi-squared association measure; (3) dependency model with PMI; and (4) dependency model with chi-squared. These non-supervised models, widely used in the literature on bracketing prediction, were applied to the whole sample of 190 MWTs.

Table 6 shows that the two predictive models, explained in this paper, outperformed the baseline models. Furthermore, the dependency model achieved better performance than the adjacency model, and the chi-squared association measure yielded better results than PMI.

Models	Precision	Recall	F_1
Adjacency model with PMI	0.6444	0.7250	0.6823
Adjacency model with chi-squared	0.6623	0.7375	0.6979
Dependency model with PMI	0.6818	0.7500	0.7143
Dependency model with chi-squared	0.7011	0.7625	0.7305
Decision tree model	0.8889	1.0000	**0.9412**
Random forest model	1.0000	1.0000	**1.0000**

Table 6: Comparison of the decision-tree and random forest models to four baseline models.

5.8 Comparison of the Models

Despite the promising results, it is obvious that further investigation is necessary to acquire a more in-depth understanding of the influence of the semantic variables in this study on ternary compound bracketing. Therefore, the following statements should be considered scope-bounded because they were derived from a restricted framework in which this research was conducted, namely specialized ternary compounds from Coastal Engineering used in sentences mentioning named rivers.

As far as the selection of the best model is concerned, there are convincing arguments in favor of either model. Since the random forest model had an error-free performance, it could be used to implement a system for bracketing ternary compounds.

Nevertheless, the performance of the decision-tree model was also fairly good. It also has the advantage of interpretability and visualization, which affords linguistic insights into how ternary compound bracketing is governed by semantic information encoded in a sentence. Since the binary decision-tree model only needed the *SemRel* predictor to achieve a highly satisfactory level of performance, practical applications for automatic bracketing could employ solely the semantic relation between a ternary compound and another argument in the same sentence.

6 Discussion

Although the comparison of our study with previous research in the literature review is far from ideal, it still serves as an indication of the performance of our semantic approach.

For bracketing prediction, previous research focused on semantic information provided by the components of an MWT. The number of variables that they used for prediction ranged from 12 to 517,254 features. These variables were mostly based on n-gram statistics, which could arguably capture some semantic information encoded in frequent co-occurrences of MWT components (Lazaridou A. et al., 2013: 1909). Other research studies relied on semantic information of the MWT components stored in linguistic resources such as WordNet. The overall accuracy of the prediction models ranged from 72.60% to 95.40%.

Our semantic approach, however, was based on semantic information that previous research has not as yet considered. The semantic information was encoded in both the co-text of a ternary compound (i.e., the sentence where the ternary compound was used as an argument) and the ternary compound seen as a unit (i.e., its semantic role). The set of variables consisted of only three (semantic relation, lexical domain, and semantic role of the MWT), whereas previous research employed a minimum of 12 variables (León-Araúz P. et al., 2021). This set of three variables yielded, in the test dataset, an error-free performance with a random forest model, whereas the highest overall accuracy achieved in previous research was 95.40% with support vector machine (Pitler E. et al., 2010), a less interpretable predictive model.

7 Conclusions

A set of 1,694 sentences, in which a named river was an argument of the predicate of the sentences, were semantically analyzed and annotated with the lexical domain of the predicates, the semantic role of the arguments, and the semantic relation between the arguments. Those semantic annotations were analyzed to see whether the bracketing of a ternary compound, when used as an argument in a sentence, can be predicted from the semantic information encoded in that sentence.

The semantic relation of the MWT to another argument in the same sentence, the lexical domain of the predicate, and the semantic role of the MWT were able to predict the bracketing of the 190 ternary compounds used as arguments in a sample of 188 semantically annotated sentences (out of the 1,694 annotated sentences). A random forest model, with three ensembled decision trees, achieved in the test dataset an AUC equal to 100% (overall accuracy of 100%). When a decision tree was trained, the model only needed the semantic relation to yield, in the test dataset, an AUC equal to 95.45% (overall accuracy of 94.74%). Hence, the semantic relation of an MWT to another argument in the same sentence proved enormous capability to predict ternary compound bracketing.

Therefore, this pilot study showed that the semantic information in a sentence, encoded in the semantic relation of the MWT to another argument in the same sentence, the lexical domain of the predicate, and the semantic role of the MWT, contributed substantially to compound parsing. Given the beneficial effects of multiword-term bracketing on overall accuracy of sentence parsers (Vadas D. and Curran J.R., 2008), and machine translation systems (Green N., 2011), this result potentially suggests a novel research direction in the integration of such semantic variables into syntactic parsers and machine translation applications, in line with Agirre E. et al. (2008), Girju R. et al. (2005), and Kim S.N. and Baldwin T. (2013).

Evidently, it is not as yet clear whether such semantic variables are also able to predict the bracketing of MWTs of four or more constituents. This issue is thus deferred for further investigation.

Finally, notwithstanding the promising results, they should be considered scope-bounded because of the small size of the MWT sample and the restricted framework in which the analysis has been conducted, namely specialized ternary compounds from Coastal Engineering used in sentences that mentioned named rivers. In future research, a wider framework shall be established to acquire a more profound understanding of the influence of the semantic variables focused in this study on multiword-term bracketing.

Acknowledgements

This research was carried out as part of projects PID2020-118369GB-I00, "Transversal Integration of Culture in a Terminological Knowledge Base on Environment" (TRANSCULTURE), funded by the Spanish Ministry of Science and Innovation; and A-HUM-600-UGR20, "Culture as Transversal Module in a Terminological Knowledge Base on the Environment" (CULTURAMA), funded by the Andalusian Ministry of Economy, Knowledge, Business, and University.

References

Agirre, E., T. Baldwin, and D. Martinez (2008). Improving parsing and PP attachment performance with sense information. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies* (pp. 317-325). ACL.

Barrière, C., and P.A. Ménard (2014). Multiword noun compound bracketing using Wikipedia. In *Proceedings of the First Workshop on Computational Approaches to Compound Analysis (ComAComA 2014)* (pp. 72-80). ACL.

Bergsma, S., E. Pitler, and D. Lin (2010). Creating robust supervised classifiers via web-scale n-gram data. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (ACL)* (pp. 865-874). ACL.

Brants, T., and A. Franz (2006). Web 1T 5-gram Version 1. Linguistic Data Consortium.

Faber, P., and R. Mairal (1999). *Constructing a Lexicon of English Verbs*. Mouton de Gruyter.

Faber, P., P. León-Araúz, and J.A. Prieto (2009). Semantic relations, dynamicity, and terminological knowledge bases. *Current Issues in Language Studies*, *1*, 1-23.

Faruqui, M., and C. Dyer (2015). Non-distributional word vector representations. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics* (pp. 464-469). ACL.

Fellbaum, C.A. (1998). Semantic network of English: The mother of all WordNets. *Computers and the Humanities*, *32*, 209-220.

Fernández, A., S. García, M. Galar, R.C. Prati, B. Krawczyk, and F. Herrera (2018). *Learning from Imbalanced Data Sets*. Springer.

Fillmore, C.J. (1968). The case for case. In E. Bach, and R. Harms (Eds.), *Universals in Linguistic Theory* (pp. 1-89). Holt, Rinehart, and Winston.

Girju, R., D.I. Moldovan, M. Tatu, and D. Antohe (2005). On the semantics of noun compounds. *Computer Speech and Language*, *19*(4), 479-496.

Green, N. (2011). Effects of noun phrase bracketing in dependency parsing and machine translation. In *49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies. Proceedings of Student Session* (pp. 69-74). ACL.

Hellmann, S., C. Stadler, J. Lehmann, and S. Auer (2009). DBpedia live extraction. In R. Meersman, T. Dillon, and P. Herrero (Eds.), *On the Move to Meaningful Internet Systems (OTM 2009)* (Vol. 5871, pp. 1209-1223). Springer. Lecture Notes in Computer Science.

James, G., D. Witten, T. Hastie, and R. Tibshirani (2015). *An Introduction to Statistical Learning*. Springer.

Kim, S.N., and T. Baldwin (2013). A lexical semantic approach to interpreting and bracketing English noun compounds. *Natural Language Engineering*, *19*(3), 385-407.

Krippendorff, K. (2012). *Content Analysis: An Introduction to its Methodology*. Sage.

Kroeger, P.R. (2005). *Analyzing Grammar: An Introduction*. Cambridge University Press.

Kuhn, M. (2021). *caret: Classification and Regression Training*. R package version 6.0-90.

Lapata, M., and F. Keller (2004). The web as a baseline: Evaluating the performance of unsupervised web-based models for a range of NLP tasks. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the ACL (HLT-NAACL 2004)* (pp. 121-128). ACL.

Lauer, M. (1994). *Conceptual Association for Compound Noun Analysis*. CoRR.

Lauer, M. (1995). Corpus statistics meet the noun compound: Some empirical results. In *Proceedings of the 33rd Annual Meeting of the ACL* (pp. 47-54). ACL.

Lazaridou, A., E.M. Vecchi, and M. Baroni (2013). Fish transporters and miracle homes: How compositional distributional semantics can help NP parsing. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing (EMNLP 2013)* (pp. 1908-1913). ACL.

Leech, G. (1981). *Semantics: The Study of Meaning*. Penguin.

León-Araúz, P., A. San Martín, and A. Reimerink (2018). The EcoLexicon English corpus as an open corpus in Sketch Engine. In *Proceedings of the 18th EURALEX International Congress* (pp. 893-901). Euralex.

León-Araúz, P., M. Cabezas-García, and P. Faber (2021). Multiword-term bracketing and representation in terminological knowledge bases.

In *Electronic Lexicography in the 21st Century. Proceedings of the eLex 2021 Conference* (pp. 139-163). Lexical Computing CZ.

Lin, D., K.W Church, H. Ji, S. Sekine, D. Yarowsky, S. Bergsma, K. Patil, E. Pitler, R. Lathbury, V. Rao, K. Dalwani, and S. Narsale (2010). New tools for web-scale n-grams. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation* (pp. 2221-2227). ELRA.

Marcus, M. (1980). *A Theory of Syntactic Recognition for Natural Language*. MIT Press.

Marcus, M.P., M.A. Marcinkiewicz, and B. Santorini (1993). Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics*, *19*(2), 313-330.

Ménard, P.A., and C. Barrière (2014). Linked open data and web corpus data for noun compound bracketing. In *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC'14)* (pp. 702-709). ELRA.

Michel, J.B., Y.K. Shen, A.P. Aiden, A. Veres, M.K. Gray, T.G.B. Team, J.P. Pickett, D. Holberg, D. Clancy, P. Norvig, J. Orwant, S. Pinker, M.A. Nowak, and E.L. Aiden (2010). Quantitative analysis of culture using millions of digitized books. *Science*, *331*(6014), 176-182.

Nakov, P., and M. Hearst (2005). Search engine statistics beyond the n-gram: Application to noun compound bracketing. In *Proceedings of the 9th Conference on Computational Natural Language Learning (CoNLL)* (pp. 17-24). ACL.

Pitler, E., S. Bergsma, D. Lin, and K.W. Church (2010). Using web-scale n-grams to improve base NP parsing performance. In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)* (pp. 886-894). ACL.

Pustejovsky, J., P. Anick, and S. Bergler (1993). Lexical semantic techniques for corpus analysis. *Computational Linguistics*, *19*(2), 331-358.

Resnik, P.S. (1993). *Selection and Information: A Class-based Approach to Lexical Relationships*. Ph.D. Thesis. University of Pennsylvania.

Roget, P.M. (1852). *Roget's Thesaurus of English Words and Phrases*. Available in Project Gutemberg. https://www.gutenberg.org/ebooks/10681.

Rojas-Garcia, J. (forthcoming). Semantic representation of context for the inclusion of named rivers in a terminological knowledge base. *Frontiers in Psychology*.

Ruppenhofer, J., M. Ellsworth, M.R.L. Petruck, C.R. Johnson, and J. Scheffczyk (2010). *FrameNet II: Extended Theory and Practice*. International Computer Science Institute.

Thompson, P., S.A. Iqbal, J. McNaught, and S. Ananiadou (2009). Construction of an annotated corpus to support biomedical information extraction. *BMC Bioinformatics*, *10*, 349.

Vadas, D., and J.R Curran (2007). Large-scale supervised models for noun phrase bracketing. In *Proceedings of the 10th Conference of the Pacific Association for Computational Linguistics (PACLING-2007)* (pp. 104-112). PACLING.

Vadas, D., and J.R. Curran (2008). Parsing noun phrase structure with CCG. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies* (pp. 335-343). ACL.