AUTOMATIC DISCOVERY OF MULTIWORD NOUNS BASED ON SYNTACTIC-SEMANTIC REPRESENTATIONS

Xiaoqin Hu

Beijing Language and Culture University, Beijing, China

ABSTRACT

This research aims to explore a deeper representation of the internal structure and semantic relationship of multiword nouns (MWNs) for improving MWN discovery. This representation focuses on MWN formations, which follow a series of categorical and semantic constraints. Linguistically motivated semantic features are defined by computing the internal semantic relations of MWNs. The internal structures are represented by describing categorical combinations in a hierarchy, and the internal semantic relations are represented with the help of semantic combinations of constituents. The results show that combining linguistically motivated semantic features with statistically motivated semantic features improves MWN discovery.

KEYWORDS

automatic discovery of multiword nouns, internal structure and semantic relation, categorical and semantic constraints, linguistic knowledge

1. INTRODUCTION

Multiword nouns (MWNs) are a specific category of multiword expressions (MWEs) that consist of multiple lexemes but display lexical, syntactic, semantic, pragmatic and/or statistical idiomaticity (Baldwin and Kim, 2010). Automatic MWN discovery is an MWE discovery task focusing only on a specific category and aims to discover new MWNs (types) in text corpora, different from MWE identification, which is the process of automatically annotating MWEs (tokens) in running text by associating them with known MWEs (types) (Constant et al., 2017). Automatic MWN discovery is always a challenging issue in NLP due to the prevalence and productivity of MWNs. Many studies apply joint parsing models, machine translation or part-of-speech (POS) tagging and MWE processing (MWE identification and MWE discovery) to improve both of them. However, the MWE characteristics are category dependent. MWNs do not exhibit discontinuity characteristics and benefit less. The research work presented in this article focuses only on MWN discovery and does not consider the support of natural language processing applications for MWN discovery.

The standard approach for automatic MWN discovery is n-gram classification based on association measures (Cavnar and Trenkle, 1994). Morphosyntactic and syntactic patterns are commonly used in preprocessing to improve candidate extraction. However, association measures are not straightforward for long chunks, especially for measures relying on contingency tables (Banerjee and Pedersen, 2003). There are also approaches that exploit the noncompositionality of MWEs for MWN discovery, such as Pearce's (2001) synonym substitution method and the semantic similarity method of Baldwin et al. (2003), which represent

David C. Wyld et al. (Eds): SOEN, SIPP, PDTCA, ITE, CCSIT, NLPCL, DaKM, BIGML, AISC -2023
pp. 197-211, 2023. CS & IT - CSCP 2023 DOI: 10.5121/csit.2023.131217

senses as vectors of cooccurring context words for discriminating MWNs by estimating the semantic similarity. However, these methods are less beneficial for many MWNs that have very low noncompositionality. Therefore, many recent studies tend to consider MWN discovery as a classification task based on statistical calculations combined with linguistic features, such as orthographic variations and capitalization, inflection, and punctuation.

Despite numerous encouraging applications for MWN discovery, most approaches are limited to the simple representation of MWNs. The benefit of adopting more complex and deeper representations capable of representing, for example, embedded MWEs (Constant et al., 2017), is still unclear. Nevertheless, many new MWNs are derived from the lexicalization of complex syntactical constructions over time, and the embedded MWNs formed from another complex construction are also quite frequent (Hu, 2019). Studies based on simple MWN representations seem less adequate and efficient for MWN discovery. A system that can exploit the deeper representations of embedded MWNs is needed.

The work presented in this article aims to exploit a deeper representation of the internal structures and semantics of MWNs to improve MWN discovery. Two types of rules, the rules describing categorical combinations and the rules describing semantic combinations in MWN formation, can be learned automatically from the lexical resource of Hu (2020), which is annotated with internal hierarchy and syntactic-semantic combinations of 3,669 noun-noun MWNs. The rules describing categorical combinations are a finite set of morphosyntactic patterns with internal hierarchical relations of MWNs and tend to be exhaustive. Then, MWN discovery is executed as a classification task by defining semantic features with the support of linguistic rules. Finally, the efficiency of linguistic semantic features defined by different grain sizes and their combination with statistical semantic features are evaluated.

2. RELATED WORK

The most popular methods for MWN discovery are the n-gram classification based on word co-occurrence measures. Pecina (2008) compared different association measures and obtained improved results by combining different collocation measures in German Adj N and PPVerb collocation discovery. Some studies also proposed weighting some very frequent words, such as prepositions, differently from regular tokens to improve association measure-based methods (Hoang et al., 2009). However, association measures are difficult to generalize to arbitrary n-gram MWN candidates. Most solutions for this issue tend to merge two-gram MWNs as single tokens and apply the measures recursively (Seretan, 2011; Da Silva et al., 1999). Nevertheless, this solution overlooks the internal MWN hierarchy and may produce more noise.

Moreover, the n-gram approach is generally limited to a finite number of grams. Sequences with more than n-grams can be missed or the accuracy is largely decreased if the defined number of grams is redundant. Many rule-based approaches have tried to take advantage of morphosyntactic descriptions. Boulaknadel et al. (2008) described French MWNs by 5 types of morphosyntactic patterns: N+A, N1+Prep+N2, N1+Prep (Det)+N2, N1+N2 and N1+Ã +Vinf for term extraction. Bourigault et al. (1996) established a series of boundary sequences, which indicate the boundaries of noun phrases in French, such as *aux* and *par une*. Biskri et al. (2004) extracted MWNs by incorporating a linguistic filter that eliminates the term candidates ending with a functional word, a verb, an adverb, or a preposition. Lebarbé (2002) also used lexicosyntactic patterns for extracting term candidates before calculating the dependency relations for term validation. However, most of these morphosyntactic descriptions are unstratified. They do not consider the internal hierarchy of MWNs and can rarely handle embedded MWNs. These morphosyntactic descriptions are difficult to exhaustively describe, and some may result in irrelevant sequences in MWN discovery.

Methods using statistical association measures for MWN discovery largely depend on the corpora profiles. Many commonly used MWNs that occur at low frequencies are easy to miss. Many recent studies have tended to exploit more linguistic knowledge to overcome this default. Rayson et al. (2004) proposed a UCREL semantic analysis system (USAS) semantic tagger that aims to assign semantic field information (including 21 major discourse fields and 232 fine-grained semantic field tags) for MWE discovery. The results were encouraging, especially for discovering low-frequency MWEs. Al-Haj and Wintner (2010) used a support vector machine (SVM) classifier, which supports morphological and morphosyntactic idiosyncratic properties to distinguish an MWE and non-MWE noun-noun construction. Their implementation achieved an F-score improvement of 1.16 and an accuracy improvement of 1.29 compared to n-gram association measure-based approaches. Gayen and Sarkar (2013) discovered Bengali MWNs based on random forest by combining statistical co-occurrence measures and syntactical features, such as word length and inflection. for noun-noun WMN discovery. Tsvetkov and Wintner (2014) defined various linguistically motivated classification features (orthographic variation, capitalization, noninsertion, etc.) for noun-noun MWN discovery. However, these methods are limited to noncomplex MWNs and require a preestablished linguistic resource depending on the language, which is sometimes not easy to obtain.

3. MULTIWORD NOUNS IN FRENCH

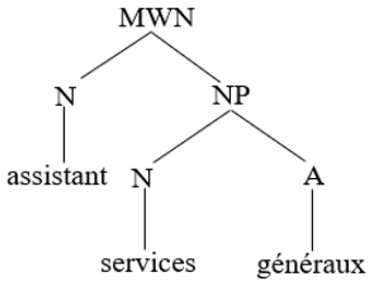
3.1. Definition of Multiword Nouns

MWNs are composed of multiple units separated by space but cannot be further analyzed by syntax, and they refer to only one semantic concept. Any complex lexical unit composed of two or more terms is considered an MWN by Guilbert (1997), Mathieu-Colas (1996), Gross (1986), Apothéloz (2002), Di Sciullo (2005), and Riegel et al. (1994). MWNs are permanent units different from nominal syntagms that are free and occasional (Grevisse, 1986). MWNs are groups of words that, taken together, have noncompositional semantics, which means that the meaning of MWNs is not made up of the sum of its parts. The introduction of the "freezing" concept distinguishes MWNs from free syntagms by applying to the structure the transformations accepted by a free syntagma of the same nature (Jacquemin, 1991).

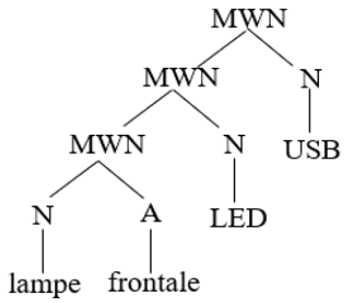
3.2. Construction of Multiword Nouns

Booij (2009) approached word formation patterns as abstract schemas that are generalized over sets of existing complex words with systematic correlations between the form and meaning. The formation process of endocentric compounds is analyzed with these schemas within the theoretical grammar construction framework (Goldberg, 2006), in which languages are considered hierarchical construction networks (Trousdale, 2013). An endocentric compound includes a head that transmits its semantic and syntactic properties to the compound (Villoing, 2012). The discussion is focused on how the complex constructions are embedded hierarchically in the nonhead position to create new compounds. Hu (2019) claimed that a complex structure could be embedded as a constituent in a pattern of an MWNs. The complex structure can be an MWN (c.f. Example (1b)) or a syntactical structure (c.f. Example (1a)). The internal hierarchical network of MWN constructions and the grammatical constraints for forming MWNs are thoroughly discussed and clarified by Hu (2020). It is argued that there exists a dependency relation between constituents of MWNs, which is the internal grammatical relation, such as attributive, subordinated, coordinated, and appositive for noun-noun MWNs.

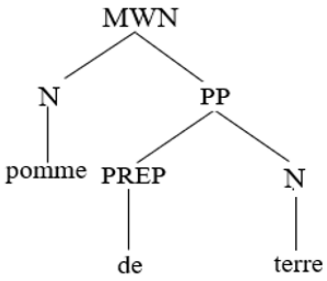
This hierarchy can be represented by a parse tree, as shown in Figure 1. Example (1a) is a representation of a two-level MWN, in which the complex structure NA *(services généraux*, 'general service') is embedded to form an MWN *assistant services généraux* ('general services assistant') (c.f. Figure 1 (a)). Then, Example (1b) presents a three-level structure, including two embeddings. This analysis is applicable to other types of MWNs. For example, *pomme de terre* ('potato'), type N+PREP+N, is annotated as N+PP (PP refers to a prepositional phrase), in which PP is a complex structure, i.e., de+N (c.f. Figure 1 (c)). Hu (2020) conducted a morphosyntactic annotation of 3,669 noun-noun MWNs (including embedded MWNs) and argued that most of the MWNs are limited to three levels and that those of four or more levels are rarely seen due to the difficulty of comprehension caused by too many levels of embedding. It is emphasized that some categorical combinations, such as structures NA and NN, are very productive in MWN formations, and the MWNs are formed within categorical constraints in the hierarchical network.



(a) MWN deeper analysis of example (1a)



(b) MWN deeper analysis of example (1b)



(c) MWN deeper analysis of example (1c)

Figure 1. Hierarchy of the internal structure of multiword nouns

3.3. Semantics of Multiword Nouns

Internal semantic relations are associated with grammatical relations. Hu (2020) provided a thorough description of the distribution of semantic relations between N1 and N2. For example, two constitutive elements cooccurring with a coordinated relation are two paronyms; those with an appositive relation are generally hyponymies; a subordinated relation often implies a semantic relation, such as actant, purpose, manner, location or time; and an attributive relation contains semantic relations, such as form, style, function, etc. Semantic relations reflect potential predicates between constituent nouns and can be represented by semantic classes of constituents. The semantic class is defined as a set of lexical units that share a series of common syntactic-semantic properties (Gaston, 1994); for instance, *book*, *article*, *novel* and *essay* belong to the semantic class called <text>. The nouns of <text> share the same semantic property; they refer to the objects that are written and have the same syntactic behavior. The predicate *edit* or *write* take the nouns of <text> as arguments.

Hu (2020) defined five coarse-grained semantic relations of MWNs, such as PROPERTY, ADVERBIAL USE, PARANYMY, ACTANT and BE, and approximately forty fine-grained semantic relations, such as DO, CONTENT, LIKE, OF, NAME, and FOR. The semantic relation PROPERTY specifies a particular object or being and normally groups semantic class combinations, such as clothes+material (e.g., *gant saxe*, 'saxon glove'), furniture+ style (e.g., *étagère design*, 'shelf with a sense of design'), and appliance+part (e.g., *parapluie canne*, 'umbrella cane'). The semantic relation BE generally groups semantic combinations, such as *île Maurice* ('island of Mauritius') and *médecin généraliste* ('general practitioner'), in which N_1 (e.g., *Maurice*, 'Mauritius') is the name of N_2 (e.g., *île*, 'island') or N_1 (e.g., *médecin*, 'doctor') is the hypernym of N_2 (e.g., *généraliste*, 'general'). The semantic relation PARANYMY generally groups semantic combinations, such as profession+profession (e.g., *boulanger pâtissier*, 'pastry baker'), clothes+clothes (e.g., *jupe culotte*, 'underwear skirt'), and chemical substance+chemical substance (e.g., benzène sulfamide, 'benzene sulfonamide'). Therefore, it can be assumed that the semantic class combination reflects, to some extent, the semantic relation between two constituents.

4. RESOURCES

To assess the performance of the proposed method for MWN discovery, especially embedded MWNs, a corpus that contains a large number of MWNs with a great variety of types is needed. The corpus was built from the web using the RENE corpus building tool (Hu, 2015), which collected only the linguistically valuable web content from given websites. Different genres of texts were selected (news, guides, blogs, tips, cookbooks, and employment ads) concerning various subjects from web communities, such as Comment ça marche, Cnet France, and Doctissimo. These web texts covered different themes and various types of MWNs and included many new MWNs that may not be registered in dictionaries but may also include several syntactic or orthographic errors. The established corpus includes a total of 11,274 sentences, 453,650 tokens, and 189,219 words covering various kinds of vocabularies. The corpus was clean and did not contain duplicates.

The corpus was built from the web using the RENE corpus building tool (Hu 2015), which collects only the linguistically valuable web content from given websites. Different genres of texts were selected (news, guides, blogs, tips, cookbooks, and employment ads) concerning various subjects from web communities, such as Comment ça Marche, Cnet France, Meilleur du Chef, Geek Food, Auto Cara, Doctissimo, Internaute, Futura, Au Féminin, Ciao, and Blogautomobile. It includes a total of 11,274 sentences, 453,650 tokens, and 189,219 words covering various kinds of vocabularies (artifact nouns, occupational nouns, scientific nouns, and other general nouns). The corpus was clean and did not contain duplicates.

The corpus was preprocessed by the Stanford log-linear part-of-speech tagger for tokenization and POS tagging (Toutanova and Manning 2000; Toutanova et al. 2003). The Stanford log-linear POS tagger is an entropy-based POS tagger that is used via a dependency network representation and uses both preceding and following tag contexts, fine-grained modeling of unknown word features, and rich lexical features, such as features for disambiguating verb tense. However, the Stanford log-linear POS tagger does not allow access to the lemmas of words. Thus, DELA (Paumier et al. 2003; Courtois and Silberztein 1990) was used for lemmatization. DELA is distributed with Unitex by using DELA syntax, which describes the simple and compound lexical entries of a language with their grammatical, semantic and inflectional information. The lexical resource DELA is available for multiple languages. The French DELA contains 683,824 single-word entries corresponding to 102,073 lemmas and 108,436 multiword entries corresponding to 83,604 MWEs.

To generate linguistic rules, the lexical resource of compound nouns of Hu (2020) needs to be explored. It is an annotated compound noun resource expanded from DELA and includes 13,955 artifact compound nouns, 3,735 occupational compound nouns, 3,992 scientific compound nouns and 46,471 general compound nouns, for a total of 68,153 compound nouns. Each entry is annotated with four levels of linguistic information: the internal structure, compounding method combination, internal semantic relation and semantic class combinations. All linguistic information is annotated considering the internal hierarchy of compound nouns. However, only categorical and semantic constraints are explored in our research. The internal structure is annotated in the hierarchical network, as shown in Example (2),

(2) (NATC (NATC (V ouvre)-(N boîte)) (NATC (V affute)-(N couteaux)))
open+box+sharpen+knife 'bottle opener and knife sharpener'

The nodes indicate the compounding method (such as NATC, which refers to native compounding) or grammatical category (such as V and N), and the internal hierarchy is marked by brackets. Semantic relation and semantic combination information are annotated in square brackets following each entry, as shown in Example (3),

(3) NATC (N wagon) - (N fumeur) [SUBO,vehicle container+person,R=FOR,
REF=wagon for smoker]

SUBO (subordinated) indicates the grammatical relation between two constituents, and R refers to the semantic relation. "Vehicle container+person" is the semantic class combination of the two constituents. The semantic information of the embedded MWNs is no longer annotated in other entries, but their semantic information is annotated when they are registered as independent entries.

Wordnet Libre du Français (WOLF, 'Free French WordNet') is also required for semantic category annotation, which is a free semantic lexical resource (WordNet) for French built from Princeton WordNet (Miller et al., 1990) and various multilingual resources (Sagot and Fišer, 2008) for automatic translation, extension, and manual validation tasks.

5. GENERATION OF LINGUISTIC RULES

5.1. Rules of Categorical Constraints

According to Hu (2020), compounds of more than three levels are extremely rare. It is argued that the embedding iteration is not infinite since too many embedding levels hinder the comprehension of interlocutors. Therefore, only the patterns of one to three levels are considered in the processing. This ensures recall and avoids the noise introduced by too many iterations. Different patterns of noun-noun compounds are extracted from the lexical resource of compound nouns expanded by Hu (2020). Each of the extracted patterns is interpreted as a sequence of grammatical codes, as shown in the first column of Table 1. The internal hierarchy is indicated by the index numbers of the grammatical codes in the sequence. A colon ":" refers to the highest level, a hyphen "-" implies a lower level, and a tilde "∼" indicates the lowest level in the hierarchy of compound nouns. The same sequence of codes may have several interpretations of the internal hierarchy; thus, they are enumerated as different structures in the rules.

Table 1. Rules of categorical constraints

CODES	HIERA-CHY	PATTERN	CODES	HIERACHY	PATTERN
N N	0:1	N+N	V:pres N V:pres N	0-1:2-3	VN+VN
N N N	0:1-2	N+NN	N ADV ADJ N	0-1-2:3	NADVA+N
N N N	0-1:2	NN+N	N N N ADJ	0-1:2-3	NN+NA
N N ADJ	0:1-2	N+NA	N ADJ N N	0-1:2-3	NA+NN
N V:pres N	0:1-2	N+VN	N ADJ N V:ppre	0-1:2-3	NA+NVpp
N N V:ppre	0:1-2	N+NVpr	N PRP N N	0-1-2:3	NPrepN+N
N N V:ppre	0:1-2	N+NVpp	N PRP N N N	0-1-2:3-4	NPrepN+NN
N ADJ N	0:1-2	N+NP	N ADV ADJ N N	0-1-2:3-4	NADVA+NN
N NUM N	0:1-2	N+NP	N N N PRP N	0-1:2-3-4	NN+NPrepN
V:pres N N	0-1:2	VN+N	N N N N	0:1-2∼3	N+N+NN
N ADJ N	0:1-2	NA+N N	ADJ N N	0∼1-2:3	NA+N+N
N N PRP N	0:1-2-3	N+NPrepN	N N N ADJ	0:1-2∼3	N+N+NA
N N N N	0-1:2-3	NN+NN	N N N N	0-1∼2:3	N+NN+N
N N N ADJ	0-1:2-3	NN+NA	N N NUM N	0:1-2∼3	N+N+NP
N N ADJ N	0-1:2-3	NN+NP	N N PRP N PRP N	0:1-2-3∼4∼5	N+N+P+NPrepN
N N NUM N	0-1:2-3	NN+NP			

5.2. Rules of semantic constraints

The rules of semantic constraints consist of two parts: semantic category resources and semantic relation seeds. The semantic class of one word is considered its semantic category (c.f. section 3.3). The semantic category resource is generated automatically from WOLF. The semantic category resource indicates the semantic category of each entry and allows for providing a semantic tag reference. The semantic relation seeds are generated automatically from the lexical resource of Hu (2020). The latter is first annotated with the semantic category resource expanded from WOLF, from which the semantic relation seeds are then automatically learned.

5.2.1. Semantic category resource

The construction of semantic category resources presented in this article considers different levels of semantic classes. The hypernyms are taken as semantic classes of the entries in WOLF. For polysemous words, only the most frequent meaning is adopted. The closest hypernym of one entry in WordNet is defined as a direct hypernym, and the others are all inherited hypernyms. For example, the direct hypernym of *fork*: *cutlery* and all other inherited hypernyms linked to *fork* given in WordNet are cited in Figure 2. From cutlery to entity, the semantics of hypernyms become increasingly general. Thus, direct hypernyms are considered the most fine-grained semantic classes, and conversely, the farthest inherited hypernyms are considered the coarsest-grained semantic classes. The hypernyms of each entry are captured successively from the direct entry to the farthest entry in WOLF. However, there may exist a dislocation between the captured semantic categories of two entries of hypernymy. As shown in Figure 2, the entry *cutlery* is the direct hypernym of the *fork*, and the entry *tableware* is a farther hypernym of the *cutlery*. The inherited hypernyms of *cutlery* thus constitute a subset of the inherited hypernyms of the *fork*, and the inherited hypernyms of *tableware* constitute a subset of the inherited hypernyms of *cutlery*. In addition, the entries of a different semantic nature having different sets of hypernyms could also cause dislocation (c.f. Figure 3). It can be understood that the depth of the inherited hypernyms varies with different entries.

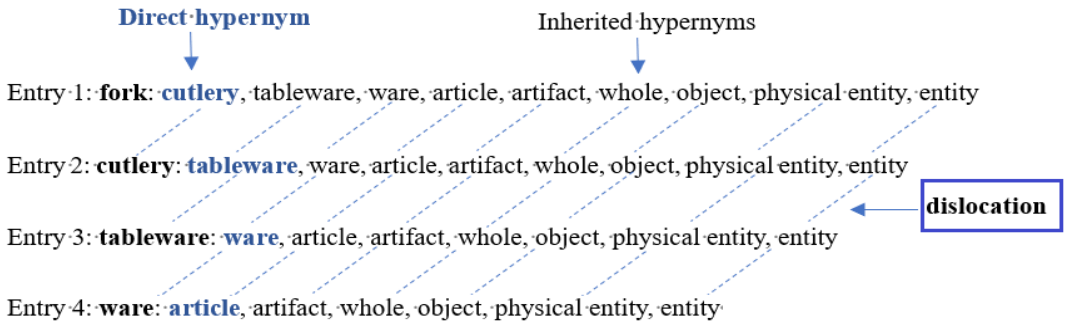


Figure 2. Dislocation of semantic categories between hypernyms

Figure 3. Dislocation of semantic categories between entries of different semantic natures

To solve this problem, the following algorithm is adopted for capturing semantic categories:

1) the longest depth of hypernyms in WOLF is first found and is used as a reference;
2) for those whose hypernym depths are inferior to the reference depth, their direct hypernyms (the most fine-grained class) will be duplicated some number of times to fill the vacancies to align with the reference entry;
3) for the entries that do not possess any hypernyms (neither a direct hypernym nor an inherited hypernym), the entry is considered its semantic class, and the vacancies are filled by duplicating the entry to align with the reference entry.

Figure 4 gives an example that shows the solution to the problem of dislocation. This semantic capturing is limited to nouns, adjectives and verbs since prepositions are considered nonsemantic value units.

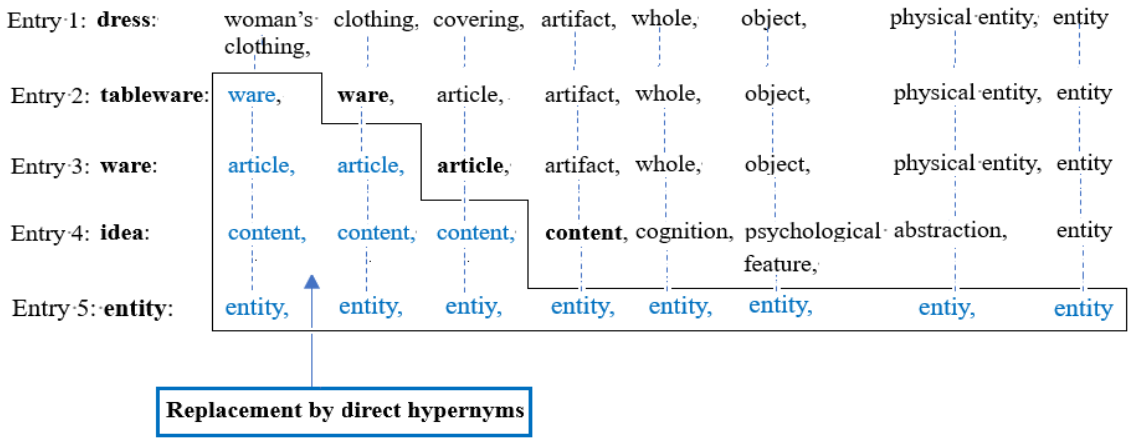


Figure 4. Solution to the dislocation problem

5.2.2. Semantic relation seeds

The semantic category of constituents of each entry in the lexical resource of Hu (2020) is first annotated with the support of the expanded semantic category resource (c.f. 5.2.1.). Then, each entry is associated with some levels of semantic class combinations according to the internal hierarchy extracted (c.f. Figure 5). Each semantic relation is represented by the combination of a pair of semantic classes.

The semantic category is annotated differently according to the types of embedded constituents of one MWN. Table 2 shows the rules of semantic category annotation for different types of constituents. When the constituent of MWN is another MWN of type N1N2, N1A or N1PrepN2, the semantic class of the constituent is considered the equivalent to the semantic class of N1 since most of the formed noun-noun or noun-modifier compounds specify a subclass of the head noun (Hu, 2020). For the constituents that are MWNs of VN, <artifact> ('artifact') is taken as the finest semantic class. If the constituent is a noun phrase, the semantic class of the head noun of the noun phrase is considered the equivalent to the semantic class of this constituent. <nom> ('name') is defined as the finest semantic class for the embedded constituents, which are proper nouns (whether known or unknown). For the constituents that are unknown common nouns, their semantic category annotations are annotated by referring to the semantic class of their synonyms or paronyms defined in WOLF.

In the expanded semantic relation seeds, each entry is associated with levels of combinations of two semantic classes. For a single-level MWN, its semantic relation is represented by the semantic class combination of its two constituents. For multilevel MWNs, semantic relation capture starts from the lowest level of the MWN (c.f. Figure 5). However, the embedded complex structure noun phrase is considered a single unit on the lowest level, and semantic capturing is no longer continued in their internal hierarchy.

6. METHOD

The monolingual corpus is first preprocessed, including tokenization, POS tagging and lemmatization. Then, the candidates and the internal hierarchy of each candidate are extracted based on the rules of categorical constraints. Third, linguistically motivated semantic features are generated with the support of the established rules of semantic constraints. The vector of semantic features of each candidate is a set of semantic class combinations of different grain sizes and is structured according to the internal hierarchy of the extracted candidate. The statistical

semantic features are equally generated based on the lexical association measures. Finally, sets of positive and negative samples are annotated by referring to the lexical resource of (Hu, 2020), which is partially annotated by humans, and the classifiers are trained with WEKA (Hall et al., 2009) for the experiments.

6.1. Preprocessing and candidate extraction

The corpus is parsed and tokenized by Stanford PosTagger ((Toutanova et al., 2003)), and only the POS tags are taken in this step. The lemmatization is then applied with the support of the French dictionary DELA. The candidate extraction starts from the single-level structures (without embedding) to the three-level structures (with two levels of embedding) (c.f. 3.2). Each identified sequence, whether it occurs as an independent lexical unit or is an embedded constituent of another lexical unit in the corpus, is registered as a candidate. The sequences identified by the same patterns but with different internal hierarchies are registered as different candidates.

6.2. Generation of semantic features

6.2.1. Linguistically motivated semantic features

The semantic category of each candidate is annotated with the support of the established semantic category resource and semantic category annotation rules (cf. Table 2). The semantic annotation starts from the lowest level. The semantic category of one candidate consists of semantic class combinations on different levels (from the lowest level to the highest level), and the semantic combinations on each level are a finite set of semantic combinations generated with different grain-size semantic classes (from the most fine-grained one to the most coarse-grained one) (c.f. Figure 5).

Table 2. Rules of semantic category annotations

Types of constituents	Semantic classes
Simple word N	Semantic class of N
Multiword nouns of N1N2, N1A or N1PrepN2	Equivalent to the semantic class of N1
Multiword nouns VN	“artifact”+hypernyms of “artifact”
Noun phrases	Equivalent to the semantic class of a head-noun
Proper nouns	“ nom”+hypernyms of “nom”
Unknown common nouns	Equivalent to the semantic class of synonyms/paronyms

The value assignment takes the established semantic relation seeds (c.f. 5.2.2) as a reference. For each level Li ($1 \leq i \leq k$), there exists a finite set of standard features Sj ($1 \leq j \leq m$). Each feature is a different grain-size semantic class combination Cij and Sj = ($C1j$, $C2j$., Cij..., Ckj). These combinations are searched in the semantic relation seeds. If a certain grain-size semantic combination on a certain level, denoted as Cij, is found, the value is assigned a score of 1; otherwise, it is assigned a score of 0. If the value is 0, the search is stopped, and 0 is defined as the final value of Sj. If the value is not 0, the search is continued on $Li+1$, and the value is replaced by the newly obtained value. The assignment rules are applied recursively until the end. Finally, each candidate is defined by m values V = ($v1$, $v2$..., vj..., vm), each of which is 0 or 1. Figure 5 gives an example of a feature vector. The *lampe frontale LED USB* is a three-level MWN, including two levels of embedding (c.f. Example (1b) in Figure 1). On the lowest level (L1) are the different grain-size semantic class combinations of *lampe* and *LED* (from the most fine-grained $S1$ to the most coarse-grained $S8$); and on the higher level (L2) are the semantic class combinations of *lampe LED* and *USB* (c.f. Table 2).

	S_1	S_2	S_3	S_4	S_5	S_6	S_7	S_8
L_1	C_{11} (appareil+diode)	C_{12} (appareil+semi-conducteur)	C_{13} (appareil+chef d'orchestre)	C_{14} (appareil+ appareil)	C_{15} (artefact+ artefact)	C_{16} (totalité+ totalité)	C_{17} (objet+ objet)	C_{18} (entité+ entité)
L_2	C_{21} (appareil+ périphérique informatique)	C_{22} (appareil+ périphérique informatique)	C_{23} (appareil+ périphérique informatique)	C_{24} (appareil+ équipement électronique)	C_{25} (artefact+ artefact)	C_{26} (totalité+ totalité)	C_{27} (objet+ objet)	C_{28} (entité+ entité)

Figure 5. Example of the semantic feature structure of *lampe LED USB*

6.2.2. Statistically Motivated Semantic Features

In our research, seven standard statistical features are defined for the experiments: the log-likelihood (Dunning, 1993), pointwise mutual information (Church and Hanks, 1990), the T-score, the phi coefficient (Liebetrau, 1986), the chi coefficient (Bell, 1962), salience (Kilgarrif and Rosenzweig, 2000), and Poisson stirring (Walsh, 1995). As the embedded structures are not necessarily MWNs, they can probably have a syntactical construction (a free construction, for instance, a noun phrase), and it seems inefficient to apply the association measure recursively by merging two-word sequences on each level due to the high association in the previous pass, as proposed by Seretan (2011). In our method, only the association measure on the highest level is calculated, and the lower-level words are merged as co-occurrence sequences for association calculation.

6.3. Training and Classification

For evaluation, sets of positive and negative instances were annotated. From the expanded corpus, 4,544 candidates were extracted in the candidate extraction step, in which MWNs were first annotated according to the lexical resources of Hu (2020), and then human annotations are performed to expand the list of positive instances. Three French-speaking annotators were asked to annotate the MWNs. Only the instances for which all three agreed were included. Finally, 1,827 instances were tagged as MWNs, and 2,717 were tagged as non-MWNs. The guidelines for noun-noun MWN annotation are given as follows:

• Noncompositional meaning. An MWN may accept a certain substitutability of elements, but the meaning has limited references. For example, there are certain substitutes for the second element of vin rouge ("wine+red"), such as vin blanc ("wine+white") and vin rosé ("wine+pink); the second elements, which are red, white and pink, do not represent wine color but a type of wine with a predominant color derived during fermentation from the natural pigment in the skins of grapes.
• Single referent. The meaning of some MWNs is compositional, but they always function as a single word and have a single referent. For example, the meaning of format A4 is compositional, which means a format that is A4, but it occurs as a single word and only has a single referent, which refers to a size.
• Morphological property. If an expression occurs with a connecting hyphen, it could very likely be an MWN.
Then, each candidate is associated with a vector containing the values of the defined features with their interhierarchical relations. All the classifiers offered by WEKA are trained for experiments to determine high-performance classifiers and to obtain better results. As a general method of evaluation, 10-fold cross-validation evaluation is performed by using the training materials.

7. EXPERIMENTS AND RESULTS

7.1. Experiments based on linguistically motivated semantic features

MWN discovery was first conducted based on linguistically motivated semantic features. The extracted MWN candidates were tagged with semantic categories defined by different grain sizes. Fourteen levels (grain sizes) of semantic categories were extracted from WOLF, and the results obtained by each level were evaluated. For the result of each level, the candidates tagged with a score of "1" were grouped as MWNs, and others were taken as non-MWNs. The 14 experimental results are shown in Table 3. The semantic category tagging from levels 1 to 5 obtained the highest precision (0.636) but also the lowest recall (0.450). With the coarsening of the granularity, the precision decreases, and the recall increases concurrently. However, the precision tended to be stable and decreased slowly when the semantic category was defined at levels 7 and 8. The F-score reached the highest when the semantic category was defined at level 12, and then, from level 13, the F-score decreased slowly. Most false positives were introduced by POS tagging errors.

Table 3. MWN discovery evaluation per level of semantic category defined by different grain sizes

	Recall	Precision	F-score
Level-1	0.450	0.636	0.527
Level-2	0.450	0.636	0.527
Level	0.450	0.636	0.527
Level-4	0.450	0.636	0.527
Level-5	0.450	0.636	0.527
Level-6	0.450	0.635	0.527
Level-7	0.453	0.634	0.528
Level-8	0.455	0.634	0.530
Level-9	0.482	0.600	0.535
Level-10	0.532	0.577	0.554
Level-11	0.773	0.479	0.592
Level-12	0.846	0.461	0.597
Level-13	0.878	0.448	0.593
Level-14	0.880	0.446	0.592

Second, the 14 levels of semantic categories were all applied as semantic features of each candidate for the experiment. The results obtained by IBK, random committee (RC), random tree, random forest, multilayer perceptron (MP), voted perceptron (VP) and Bayes net were superior to other classifiers offered by WEKA (c.f. Table 4). The IBK, RC and random tree classifiers permitted the highest recall of 0.462 and the highest F-score of 0.534. The highest precision of 0.636 was obtained with VP, but its recall of 0.452 was much lower. The result obtained by combining all grain sizes of semantic categories yielded little improvement. Although the highest F-score (0.534) obtained by combining the overall grain sizes of semantic categories was inferior to the highest F-score (0.597 at level 12) obtained by defining a single-level semantic category, the recall and the F-score seem slightly improved when the precision was superior to 0.600.

Table 4. MWN discovery evaluation based on the combination of the overall semantic categories

	Recall	Precision	F-score
IBK	0.462	0.633	0.534
RC	0.462	0.633	0.534
Random Tree	0.462	0.633	0.534
Random Forest	0.461	0.632	0.534
MP	0.458	0.633	0.531
VP	0.451	0.636	0.528
Bayes Net	0.452	0.635	0.528

7.2. Experiments based on statistically motivated semantic features

Experiments on MWN discovery based on association measures were also completed on the expanded corpus. The defined association measures were combined for binary classification. Then, the results obtained by combining different association measures were evaluated (c.f. Table 5). The naive Bayes multinomial (NBM), naive Bayes multinomial updateable (NBM update), random committee (RC), IBK, random tree, Bayes net and random forest classifiers obtained better results than other classifiers in WEKA. NBM obtained the highest recall of 0.367 and the highest F-score of 0.415. The highest precision was obtained with PART (0.714), but this classifier presented the poorest performance in recall rate (0.06) among all the classifiers offered by WEKA. Compared to the results obtained based on linguistically motivated semantic features, the results obtained by association measure-based methods were much poorer.

Table 5. Results of experiments based on the combination of overall defined association measures

	Recall	Precision	F-score
NBM	0.367	0.468	0.415
NBM Updateable	0.358	0.471	0.407
RC	0.262	0.459	0.334
IBK	0.260	0.460	0.332
Random Tree	0.258	0.465	0.332
Bayes Net	0.242	0.501	0.326
Random Forest	0.242	0.469	0.320
PART	0.06	0.714	0.110

7.3. Combined Linguistic and Statistical Feature Experiments

The linguistic and statistical features were finally combined for the experiment: a total of 14 different-grained linguistic features and seven statistical features. The results of the experiments are shown in Table 7. Compared to the semantic category-based methods and the association measure-based methods, the combination of both linguistic and statistical features yielded a substantial improvement. Although the precisions obtained by some classifiers decreased slightly compared with the results obtained by combining the overall semantic categories, the recall and the F-score improved by combining both the linguistic and statistical features. The highest precision reached 0.664 with the filtered classifier (FC), but the recall rate was lower. Bayes net

and naive Bayes obtained recall rates and precision rates superior to those obtained based only on semantic categories or statistical association measures.

Table 6. Results of combined linguistic and statistical experiments

	Recall	Precision	F-score
Random Committee	0.561	0.556	0.558
IBK	0.559	0.555	0.557
Random Forest	0.543	0.570	0.556
Random Tree	0.551	0.559	0.555
Bagging	0.549	0.616	0.549
J48	0.482	0.635	0.548
Random Subspace	0.470	0.648	0.544
Logistic	0.467	0.633	0.537
Naive Bayes	0.466	0.631	0.536
Bayes Net	0.460	0.632	0.533
Filtered Classifier	0.442	0.664	0.531

8. CONCLUSIONS

The present research aims to explore the internal structure and semantic relation of MWNs for improving MWN discovery. It is considered that a complex structure, an MWN or a syntactical structure could be embedded to form MWNs, and there can be 1-3 levels of embedding. The extraction of MWN candidates and linguistically motivated semantic features are based on the rules of categorical constraints and the rules of semantic constraints, respectively. Both linguistic rules are automatically generated from the pre-expanded knowledge resource. The linguistically motivated semantic features are different grain sizes of semantic class combinations. The fine-grained semantic classes tend to be hypercritical in accepting a candidate as an MWN, while the coarse-grained semantic classes seem too greedy to discriminate MWNs. The combination of linguistically motivated semantic features and statistically motivated semantic features shows a significant improvement and is superior to both semantic category-based methods and association measure-based methods. The presented methods based on linguistic knowledge of the MWN formation process are much less dependent on the corpus profile and more efficient for discovering multiple words of complex structures. The exploration of other types of features, such as features of MWN idiosyncratic behavior, can be the focus of future research.

ACKNOWLEDGMENTS

This research project is supported by the Science Foundation of Beijing Language and Culture University (supported by "the Fundamental Research Funds for the Central Universities") (Approval number: 17YBB14).

REFERENCES

[1] Al-Haj, H. & Wintner S., (2010) "Identifying Multi-word Expressions by Leveraging Morphological and Syntactic Idiosyncrasy", in *Proceedings of the conference COLING 2010*. Beijing, China.

[2] Apotheloz, D., (2002) *La construction du lexique français, Principes de morphologie derivationnelle*. Paris, Ophrys.

[3] Baldwin, T., Bannard C., Tanaka T. & Widdows D., (2003) "An empirical model of multiword expression decomposability", in *Proceedings of the ACL 2003 Workshop on MWEs*, pp89–96, Sapporo.

[4] Baldwin, T., & Kim, S. N., (2010) "Multiword expressions", in N. Indurkhya and F. J. Damerau (ed.), *Handbook of Natural Language Processing*, chapter 12, pp267–93. CRC Press.
[5] Banerjee, S. & Pedersen, T., (2003) "The design implementation, and use of the ngram statistics package", *Proceedings of CICLing 2003*, Mexico City, pp370–381.
[6] Bell, C. B., (1962) "*Mutual information and maximal correlation as measures of dependence*", Ann. Math. Statist., Vol. 33, pp587–95.
[7] Biskri, I., Meunier J.-G., & Joyal S., (2004) "L'extraction des termes complexes: une approche modulaire semi-automatique", *7es Journées internationales d'Analyse statistique des Données Textuelle*, France, pp192-201.
[8] Booij, G., (2009) "Lexical integrity as a formal universal: A constructionist view", in S. Scalise, E. Magni and A. Bisetto (ed.), *Universals of language today*, Dordrecht, Springer, pp83–100.
[9] Boulaknadel, S., Daille B. & Aboutajdine D., (2008) "Acabit : un outil d'extraction des termes complexes", http://tal.ircam.ma/conference/docs/ticam2008/ 6 boulaknadel.pdf, visité le 9 juillet 2013.
[10] Bourigault, D., Gonzalez-Mullier I. & Gros C., (1996) "LEXTER: A Natural Language Tool for Terminology Extraction", *Proceedings of the 7th EURALEX International Congress*, Goteborg, pp771–79. ¨
[11] Cavnar, W. B. & Trenkle J. M., (1994) "N-gram-based Text Categorization", *Proceedings of SDAIR-94*, Las Vegas, Nevada, U.S.A., UNLV Publications/Reprographics, pp161–71.
[12] Chakraborty, T., (2010) "Identification of Noun-Noun(N-N) Collocations as Multi-Word Expressions in Bengali Corpus", *Proceedings of 8th International Conference on Natural Language Processing (ICON 2010)*.
[13] Chang, B., Pernilla D. & Wolfgang T., (2002) "Extraction of translation unit from Chinese-English parallel corpora", *Proceedings of the First SIGHAN Workshop on Chinese Language Processing*, Morristown, NJ, pp1–5.
[14] Church, K. & Hanks P., (1990) "Word association norms, mutual information, and lexicography", *Computational Linguistics*, Vol. 16, No. 1.
[15] Constant, M., Sigogne A. & Watrin .P,. (2012) "La reconnaissance des mots composés à l'é preuve de l'analyse syntaxique et vice-versa : évaluation de deux stratégies discriminantes", *Actes de la 19e conférence sur le Traitement Automatique des Langues Naturelles*, Grenoble, pp57-70.
[16] Constant, M., Le Roux J. & Tomeh N., (2016) "Multilingual deep lexical acquisition for HPSGs via supertagging", *Proceedings of NAACL/HLT 2016*, San Diego, CA, pp1095–1101.
[17] Constant, M., Eryigit G., Monti J. & Van der Plas L., (2017) "Multiword Expression Processing: A Survey", *Computational Linguistics*, Vol. 43, No. 4, pp837-92.
[18] Courtois, M. & Silberztein M., (1990) "Le dictionnaire électronique DELAC", *Langue française*, Larousse, Vol. 87, pp71-83.
[19] Da Silva, J., Dias F., Guillore G. S. & Pereira Lopes J. G., (1999) "Using localmaxs algorithm for the extraction of contiguous and non-contiguous multiword lexical units", *Proceedings of the 9th Portuguese Conference on Artificial Intelligence*, London, pp113–32.
[20] Di Sciullo, A. M., (2005) "Decomposition compounds", *SKASE Journal of Theoretical linguistics*, Vol. 2, pp14-33.
[21] Dunning, T., (1993) "Accurate Method for the Statistic of Surprise and Coincidence", *Computational Linguistics. Evert, S. 2005. The Statistics of Word Co-occurrences: Word Pairs and Collocations*, Ph.D. thesis, Stuttgart: University of Stuttgart, pp61-74.
[22] Gayen, V. & Sarkar K., (2013) "Automatic Identification of Bengali Noun-Noun Compounds Using Random Forest", *Proceedings of the 9th Workshop on Multiword Expressions (MWE 2013)*, Atlanta, Georgia, pp64–72.
[23] Goldberg, A., (2006) *Constructions at Work*, Oxford, Oxford University Press.
[24] Green, S., de Marneffe M.-C. & Manning C. D., (2013) "Parsing models for identifying multiword expressions", *Computational Linguistics*, Vol.39, No. 1, pp195–227.
[25] Grevisse, M., (1986) *Le bon usage, Belgique*, Duculot.
[26] Gross, M., (1986) *Grammaire transformationnelle du français: Syntaxe du nom*. France, Cantilene.
[27] Guilbert, L., (1997) *La creativité lexicale*, Paris, Larousse.
[28] Hall, M., Frank E., Holmes G., Pfahringer B., Reutemann P. & Witten I. H., (2009) "The WEKA data mining software: An update", *SIGKDD Explorations Newsletter*, Vol. 11, pp10–18.