Boosting the Coverage of a Semantic Lexicon by Automatically Extracted Event Nominalizations

Kata Gábor[1], **Marianna Apidianaki**[1,2], **Benoît Sagot**[1], **Éric Villemonte de La Clergerie**[1]

1. Alpage, INRIA Paris-Rocquencourt & Université Paris 7, 175 rue du Chevaleret, 75013 Paris, France
2. LIMSI-CNRS, BP 133, F-91403, Orsay Cedex, France
gkata@nytud.hu, marianna@limsi.fr, benoit.sagot@inria.fr, Eric.De_La_Clergerie@inria.fr

Abstract

An important trend in recent works on lexical semantics has been the development of learning methods capable of extracting semantic information from text corpora. The majority of these methods are based on the distributional hypothesis of meaning and acquire semantic information by identifying distributional patterns in texts. In this article, we present a distributional analysis method for extracting nominalization relations from monolingual corpora. The acquisition method makes use of distributional and morphological information to select nominalization candidates. We explain how the learning is performed on a dependency annotated corpus and describe the nominalization results. Furthermore, we show how these results served to enrich an existing lexical resource, the WOLF (Wordnet Libre du Français). We present the techniques that we developed in order to integrate the new information into WOLF, based on both its structure and content. Finally, we evaluate the validity of the automatically obtained information and the correctness of its integration into the semantic resource. The method proved to be useful for boosting the coverage of WOLF and presents the advantage of filling verbal synsets, which are particularly difficult to handle due to the high level of verbal polysemy.

Keywords: lexical acquisition, nominalization, WordNet

1. Introduction

Distributional similarity has a prominent role in lexical semantics in recent years: several methods have been proposed to exploit the distributional hypothesis (Harris, 1954) and infer semantic information by identifying distributional patterns in texts. These methods offer an alternative to the manual creation of lexical resources for various applications. The majority of the work carried out in this domain aims at extracting lexical semantic properties from mono- or multilingual corpora.

Word sense induction methods generally use unsupervised clustering techniques based on contextual similarity to identify the senses of words in texts (Pantel and Lin, 2002; Véronis, 2004; Agirre and Soroa, 2007; Apidianaki and Van de Cruys, 2011; Messiant et al., 2010). However, other types of semantic information, such as hypernymy, hyponymy and other semantic relations, can also be extracted from corpora (Hearst, 1992). Relations between words belonging to different parts of speech (PoS) categories can equally be identified, e.g. derivational relations between verbs and nouns or adjectives and adverbs (Fabre and Bourigault, 2006).

This paper presents an unsupervised lexical acquisition experiment aimed at enriching a French lexical semantic resource with event nominalizations. The acquisition method makes use of distributional and morphological information to select nominalization candidates. The candidates are then used to enrich a lexical semantic resource, the WOLF (Wordnet Libre du Français (Sagot and Fišer, 2008)), created on the basis of Princeton WordNet (Fellbaum, 1998). Two complementary methods for adding new words to existing synsets will be presented, both of them exploiting the semantic relations present in WOLF. The extraction method proved to be useful for boosting the coverage of WOLF and has the capacity to fill verbal synsets, which are particularly difficult to handle due to the high level of verbal polysemy. Morpho-semantic derivational links between adjectives and adverbs have already been explored for the purpose of improving the coverage of WOLF (Sagot et al., 2009). Our work differs from earlier experiments with respect to the nature of derivational links to be identified: we do not limit our investigation to morphologically related word pairs. Another important difference is that we acquire lexical information from corpora using distributional methods instead of relying on existing lexical resources.

2. Comparison to related resources

Two French nominalization lexica are currently available: VERBACTION (Tanguy and Hathout, 2002) and Jeux de Mots (Lafourcade and Joubert, 2008). VERBACTION contains verb-nominalization pairs where the noun is produced from the verb by morphological derivation and refers to the action expressed by the verb. Nominalization candidates were extracted from lexical resources and the Internet, using the method described in (Hathout et al., 2002). In its current state, the lexicon contains 9393 verb-noun pairs. Jeux de Mots is a relational database of lexical entries collected through an online word game[1]. It contains a variety of lexical relation types (e.g. *associated term, part of, synonymy, location, characteristic*). In the interactive game, a lexical item and a relation type are shown to two users in parallel: they are asked to suggest words related to the source word by the displayed relation type. Relations suggested by both users are saved and asked to be validated by other users before being included in the lexical network. The database is constantly growing; at present, it contains

[1] http://www.jeuxdemots.org/

over 1 300 000 relation instances among which 6 752 are considered as finalized (suggested at least 25 times).
Our acquisition method presents the following advantages over currently existing resources:

- unlike VERBACTION, it is not limited to verb-noun pairs produced by morphological derivation (p.ex. : *tomber*, ('to fall') - *chute*, 'fall'; *rouler*, 'to travel/to drive' - *circulation*, 'traffic'),
- the method guarantees that verb-noun pairs are only included if they denote events,
- it is adaptable to specific domains and new corpora,
- finally, it provides a set of distributional contexts allowing to disambiguate between different senses and to position the candidates in a lexical semantic resource.

The extraction of nominalization candidates is done in three steps:

- First, semantic similarity is calculated on the basis of syntactic distribution.
- Second, a morphological module checks whether there is a morphologically related noun among the highest ranking candidates.
- Finally, the so-called '*event indicator*' score is calculated to filter out verb-noun couples that do not refer to events.

3. Acquisition of Nominalization Candidates

Event nominalizations have a special interest for language processing: they are characterized by a complex lexical structure as they often preserve the (partial or complete) subcategorization pattern of the verb they are derived from, or are semantically related to. Lexical information related to the argument structure of nouns can be exploited to improve parsing performance with respect to attachment ambiguities. Moreover, this kind of lexical information can be useful in NLP applications as diverse as information extraction or semantic role labeling (SRL).
Current works on nominalization acquisition are sparse and mostly concentrate on enriching syntactic and semantic lexica or improving semantic role labeling (Padó et al., 2008; Lapata, 2002). Annotating the argument structure of nouns with semantic roles has actually received significantly less attention compared to verbal subcategorization. As demonstrated by (Padó et al., 2008), nominal SRL is a more challenging task compared to verbal SRL, partly due to lack of data (both in terms of available lexical resources and annotated learning corpora). Nevertheless, the main reasons for the lower performance of nominal SRL are the task-specific difficulties related to verb-noun derivations: not all verbs refer to actions; not all action verbs can be nominalized; some deverbal nouns are lexicalized with a meaning which does not correspond to the event denoted by the verb.
The mapping between the arguments of verbs and their nominalized equivalents is not always straightforward and can be ambiguous. Moreover, morphologically related nominalizations are sometimes overridden by lexicalized forms blocking the use of the derived word form. Our methodology aims at identifying nominalization equivalences even in cases where there is no direct morphological link between the verb and the noun. Therefore, we talk about word pairs related by *morphosemantic* derivation.

Three types of nominalizations can be distinguished:

- action nominalizations,
- result nominalizations,
- nominalizations corresponding to a participant — most typically to the agent — of the base verb.

Result nominalizations do not correspond to an event , consequently, they usually do not preserve the argument structure of the base verb. On the other hand, nominalizations corresponding to the action itself or to one of its participants can inherit the complete subcategorization frame of the verb or a subset of it, and display it as syntactic dependents. In the present study, we are interested in action nominalizations and we hypothesize that it is possible to identify them by exploiting the distributional similarity between verbal and nominal complement structures observed in a corpus.

3.1. Corpus

The distributional similarities between nouns and verbs were calculated using information extracted from a syntactically analyzed corpus (Table 1). The corpus contains around 700 million words covering diverse topic areas and was parsed with the TAG FRMG parser with two output formats: the Passage format (used in various French evaluation campaigns) and the DepXML format. Verbal and nominal distributions were both represented as dependency triplet frequencies. A dependency triplet is composed of a pair of lemmata with a labelled dependency relation between them (2). The distribution of nouns was extracted from the Passage format of the parser output (Vilnat et al., 2010), which includes chunking and dependency information. Subcategorization frames for verbs were extracted from the DepXML format (Villemonte de la Clergerie, 2010). A full subcategorization frame indicates the complementation pattern of the verb (subject, object, oblique complements, adjuncts) together with the lemma of the head word occurring in the given position. Table 2 lists some of the syntactic contexts extracted for the noun *inauguration* and the verb *inaugurer*.

Corpus	#words	info
Wikipedia (fr)	178.9110M	encyclopedy
Wikisource (fr)	63.9771M	literature
EstRepublicain	144.8779M	press
JRC	66.5447M	EU directives
EP	41.5288M	parliamentary debates
AFP	248.3240M	400K news wires
TOTAL	744.1638M	

Table 1: Composition of the corpus

inauguration_nc	de	exposition_nc	155
inauguration_nc	de	salon_nc	125
inauguration_nc	de	musée_nc	73
inauguration_nc	de	centre_nc	69
inauguration_nc	par	président_nc	27
inauguration_nc	par	ministre_nc	24
inauguration_nc	par	roi_nc	6
inauguration_nc	par	reine_nc	4

inaugurer_v_active	subject:Reine_np	object:exposition_nc		5
inaugurer_v_active	subject:ministre_nc	object:stand_nc		3
inaugurer_v_active	subject:roi_nc	object:exposition_nc		1
inaugurer_v_active	subject:cln_cln	object:musée_nc		1
inaugurer_v_active	subject:_PERSON_m_np	object:exposition_nc	time_mod:_DATE_artf_nc	1

Table 2: Dependency triplets and subcategorization frames for *inauguration* and *inaugurer*

3.2. Distributional Similarity

The representation of verbs and nouns was constructed from the dependency triplets extracted from the parsed corpus. The reason for preferring individual dependency relations over complete subcategorization frames is twofold. First, using complete subcategorization frames could lead to a data sparsness problem; second, nouns usually do not display more than one or two of their arguments in a sentence, which makes it difficult to establish a correspondance with complete subcategorization frames.

When constructing the representation, a general filtering of dependency triplets was carried out to select the dependency relation types relevant for each verbal entry. All subcategorization pattern occurrences were extracted from the corpus and a list of generalized frames was created by deleting all lexical information from the patterns. Subsequently, the *Pointwise Mutual Information* $PMI(x\|y)$ was calculated for each pair of verb+subcategorization frame (x, y) in the corpus:

$$PMI(x\|y) = \log \frac{p\,(x,y)}{p\,(x)\,p\,(y)}, \qquad (1)$$

where the probabilities $p\,(y)$, $p\,(x)$ and $p\,(x,y)$ were calculated by maximum likelihood estimate from the corpus.

Verb-specific frame lists were then obtained by selecting the subcategorization frames with a PMI value over a threshold established during test runs. Finally, dependency relation types were extracted from these frames, yielding a list of syntactic relations characteristic of a given verb to be compared with nominal distributions. Relations included in verbal representations were not limited to arguments: adjunct functions were also considered provided that they were characteristic of the lexical entry.

Given the list of lemma-specific dependency relation types, the frequencies of relevant dependency triplets instances were included in verbal representations. No filtering was however applied when constructing nominal representations and all dependency relation types were taken into account.

Dependency triplet instances were characterized by the type of the relation, the lemma and the morphological description of the complement. The only exception were verbal complements, in which case the lexical information was replaced by the PoS category (infinitive or clause), as we consider that verbs and nouns accepting sentential arguments do not make semantic restrictions on these types of complements.

To calculate a distributional similarity from the obtained syntactic representation, an algorithm capable of establishing a mapping between verbal and nominal argument structures was needed. We observe that verbs and nouns do not realize their arguments in the same way and that some syntactic forms can be ambiguous with respect to the grammatical function:

VERB	**NOUN**
subject/passive par	de/par
object/passive sujet	de
avec	entre (e.g. *discuter, discussion*)
de/à	sur (e.g. *réfléchir,réflexion*)

The mapping between verbal and nominal dependency relations was carried out as follows. Most oblique (prepositional) complements of the verb were mapped directly to their nominal counterparts (e.g. *voter/vote au parlement, au conseil*). We resorted to heuristics to disambiguate the preposition '*de*' which, when used with event nominalizations, can correspond to the subject, the direct object or an oblique complement of the base verb with the same preposition.

In order to decide whether a given verb is transitive or not, we considered the proportion of transitive subcategorization frames over the occurrences of the verb in the parsed corpus. The verb was considered as transitive if it had at least one transitive subcategorization frame (or a frame which had the verb in passive) with a PMI value above the previously defined treshold. The preposition '*de*' was mapped to the direct object function of transitive verbs, and to the subject function of intransitive verbs. In the case of intransitive verbs, the subject of nominal candidates was used as a meta-function including both complements with the preposition '*de*' and complements with '*par*'. On the other hand, the subject function was not considered at all for transitive verbs.

We used the Dice index to calculate a distributional similarity in terms of the intersection of the lemmata occurring in the same or equivalent syntactic positions:

$$S_D(x\|y) = \frac{2\|X \bigcap Y\|}{\|X \bigcup Y\|} \quad (2)$$

To compensate the bias introduced by the generalization of complements belonging to a verbal category, a weight that was 20 times higher was given to complements with a specified lemma as opposed to complements characterized by PoS category. A minimum similarity threshold was set to rule out irrelevant candidates. For each verb, the highest ranking candidates were kept for further processing. The following modules re-rank or filter these candidates.

3.3. Morphological Similarity

The morphological similarity module detects whether there is a candidate which is likely to be produced from the verb by morphological derivation. It is important to note that this information is only taken into account if the noun is already a potential candidate, i.e. if it has a similar distribution. Morphological similarity is calculated using the edit distance and a manually constructed list of derivational suffixes. The distance is calculated between the verb stem (after deleting the infinitival suffix) and possible nominal stems, including the complete form as well as any stem obtained after matching the list of potential nominalization suffixes. The two words are considered as morphologically related if the following criteria were satisfied for at least one of the potential nominal stems:

- the edit distance is lower than 3
- the length of the verb minus the edit distance is greater than 3.

Moreover, the two forms were considered as related independently of the length of the verb if the verb and the noun only differed in the infinitival suffix.

3.4. Filtering by the Event Indicator

The third module used assigns a value to individual nominal candidates , indicating the likeliness that a noun denotes an event. This metric is equally calculated from syntactic distributions.

A recurrent type of irrelevant nominalization candidates concerns nouns exhibiting a strong distributional similarity with a verb but actually not denoting events. This occurs when a noun is accidentally characterized by the same contexts without having the same semantic relation to the lemma/lemmata in the context (e.g. *trembler*, subject: *terre* 'the earth shakes' and *pomme* de *terre* 'potato', 'earth apple'). More importantly, some nouns can have a systematically high distributional similarity with a big number of verbs due to the fact that they accept a variety of syntactic complements. Typical examples are nouns denoting quantities or behaving like determiners (e.g. *unité, kilogramme, totalité*). Finally, participant nominalizations are also likely to accept a subset of the arguments of the base verb but since they do not refer to events, they were excluded from the list of candidates. Instead of constructing a stop-word list to filter out these entries, a dynamic filtering module was used that rules out both types of false nominalizations.

The filtering is based on a value that indicates the likeliness that the noun corresponds to an event. Similarly to Arnulphy et al. (2010) who aim at extracting named entities denoting events, we relied on syntactic contexts specific to events. However, unlike in their experiment, we did not dispose of manually annotated corpora. We therefore adopted the following semi-automatical process. We defined a metric called '*event indicator*', based on the hypothesis that event nominalizations are likely to occur in argument positions with verbs that semantically select an event as their argument. This event is syntactically realized either as an event nominalization or as a complement with a verbal category (clause or infinitive). Consequently, we extracted from our corpus a list of verbs subcategorizing for infinitives or clauses.

The resulting list was manually validated. We also indicated the syntactic realization that the verb required for its nominal candidate wherever it was different from the syntactic realization of the clausal/infinitival complement (e.g. *accuser d'avoir volé - accuser de vol* with the same syntactic function, vs *refuser de signer - refuser la signature* with different syntactic functions). For every nominalization candidate, we calculated the proportion of its occurrences in these event-like syntactic contexts, compared to the total number of its occurrences. This number was then weighted by the number of different *types* of event-like contexts characterizing the noun. The threshold was defined during test runs; we opted for a value that allowed to reduce the number of candidates by 25%.

3.5. Results

The method described above was applied to our corpus to detect nominalizations for 3 351 verbs with at least 50 occurrences in the corpus. 2 424 verbs were assigned at least one nominalization; the rest of the verbs did not have any nominalization candidates for one of the following two reasons:

- the quantity of data was not sufficient to extract verb-specific subcategorization frames and create a representation of the verb's syntactic distribution,
- the verb does not denote an event (every potential candidate was filtered out by the event indicator metric).

Subsequently, we defined a reliability threshold in terms of distributional similarity. This additional step serves to eliminate further candidates of verb-noun pairs that do not denote an event. Following this filtering phase, 1 136 verbs were assigned at least one candidate. Results are presented for manual evaluation in the form of *tickets*including a suggested nominalization, the strength of the distributional similarity as a reliability metric, and a few examples of shared contexts (dependency relations and typical lexical items occurring in the dependent position) illustrating the meaning in which the verb and the noun are supposed to be related. Figure 1 illustrates nominalization tickets with the relevant syntactic contexts shared between the verb and the noun.

Among the final list of candidates, we find verb-noun couples produced by morphological derivation (e.g. *nommer*,

'to appoint' - *nomination*, 'appointment' or *protéger*, 'to protect' - *protection*), but also semantically related words without any morphological relation (e.g. *fausser*, 'to distort' - *distorsion*; *redouter*, 'to fear' - *crainte*, 'fear'). If the list contains one or more morphologically related terms, these are presented as unique candidates; otherwise, the first two candidates (ranked by distributional similarity) are presented in the ticket.

A manual evaluation was performed on 161 randomly selected tickets, among which 113 (70%) were considered as correct nominalizations. A subsequent analysis allowed to identify recurrent errors among the nominalizations suggested by our algorithm. The most frequent source of erroneous candidates was the generalization of infinitival or clausal complements in the syntactic representation: the generalization was not sufficiently compensated by the lower weight assigned to these complements and hence yielded biased representations. On the other hand, the generalization was necessary in order to avoid sparseness data problems for these verbs.

As the algorithm uses data from an automatically annotated corpus, some of the resulting errors are due to noisy input data, and in particular to wrong lemmatization in the corpus. For instance, the method suggests to connect the word *fil* ('thread', 'wire') to the verb *pendre* ('to hang'). A closer look at the contexts presented in the ticket showed that the verb 'pendre' was often mistaken in the corpus for being the lemma of the form *pendant* ('during', preposition), which corresponds to the meaning of 'fil' in the multiword expression *'au fil de'*.

It is well known that distributional methods have difficulties distinguishing words in a synonymy relation from antonyms, both relations being characterized by similar syntactic contexts. We encountered the same problem for some of our nominalization candidates (e.g. *fermer*, 'to close' - *ouverture*, 'opening').

In other cases, we obtain correct nominalization candidates but a subset of the nominal and verbal syntactic contexts cannot be mapped to each other due to different semantic roles: the two words present the same event from a different perspective (e.g. *acheter*, 'to buy' - *vente*, 'selling'). Inversely we find examples of exact nominalization with respect to the suggested semantic role, which does not refer to the same event as the verb (e.g. *enlever*, 'to kidnap' - *disparition*, 'disappearance'; *tuer*, 'to kill' - *mort*, 'death'). Finally, a set of candidates contain semantically related words which do not designate an event: *peser*, 'to weigh' - *poids*, 'weight'; *remonter*, 'to go/date back to' - *origine*, 'origin'.

Verb	Noun	Role	Example
réclamer	demande	à	*motif encontre*
		obj	*moratoire audition*
tomber	chute	sur	*voie piste*
		à	*mer place kilomètre*
		sujet	*mur météorite matière*
		dans	*escalier crevasse coma*
freiner	ralentissement	obj	*propagation croissance*
figurer	présence	sur	*feuille liste podium*
		sujet	*invité substance*
		dans	*délégation liste équipe*
rouler	circulation	sujet	*train tramway métro*
céder	recul	obj	*cent pourcent*

Figure 1: Nominalization candidates with contexts.

4. Enriching the Lexical Resource

4.1. WOLF

The nominalization candidates were used to enrich WOLF with new lexical entries. WOLF (WOrdnet Libre du Français, *Free French Wordnet*) is a freely available semantic lexical resource for French (Sagot and Fišer, 2008). It is based on and structurally equivalent to the Princeton Wordnet (PWN) version 2.0 (Fellbaum, 1998). Like any wordnet, WOLF is a lexical database in which words (lexemes, literals) are divided by parts of speech and organized into a hierarchy of nodes. Each node has a unique id, and represents a synset . A synset groups one or more synonymous words that denote the same concept.

WOLF was built automatically from the PWN 2.0 and various multilingual resources, using two complementary approaches. Polysemous lexemes were dealt with using an approach that relies on word-aligned parallel corpora in five languages, including French. Several multilingual lexica were extracted from these aligned corpora. The obtained lexica were semantically disambiguated using the wordnets of the corresponding languages. Monosemous PWN lexemes were translated by using bilingual lexica extracted from wiki resources (Wikipedia, Wiktionary) and thesauri. WOLF contains all PWN 2.0 synsets, including those for which no French lexeme is known. The latest version of WOLF includes 32 351 non-empty French synsets and 38 001 literals, and hence provides better coverage than the manually built French part of EuroWordNet (Vossen, 1999) or the automatically constructed JAWS (*Juste Another WordNet Synset*) (Mouton and de Chalendar, 2010). However, the coverage of WOLF is still limited compared to PWN (115 424 synsets for 145 627 literals). The work presented in the following sections is aimed at increasing the coverage of WOLF by automatically assigning lexical entries to verbal synsets.

4.2. Heuristic Method

New entries can be added to WOLF either by filling empty synsets or by adding synonyms to non-empty synsets. As a first attempt, we started filling up the verbal synsets. Incidentally, they are the most difficult to handle due to the high level of polysemy that characterizes many verbs.

The synsets to be filled are selected by exploiting the derivational links already present in WOLF (coming from the structure of the Princeton WordNet). For each pair of a verb and a nominalization candidate, if the noun figured in some WOLF synset(s), we extracted the verbal synsets linked to it (them) by a derivational relation. We obtained candidates of the form {V,S}, meaning that verb V is suggested to be added to synset S. 2 353 candidates were created in this way.

The first approach used to add verbs to empty synsets was

Verbe v	**Synset** s	s **in WOLF**	**definition in Princeton WordNet**	$\mathrm{weight}(v, s)$
diversifier	ENG20-00424026-v	diversifier	vary in order to spread risk or to expand	0.88
collecter	ENG20-02238144-v	collecter, recueillir, réunir	get or gather together	0.88
collecter	ENG20-01340552-v	recueillir, collecter	assemble or get together	0.88
collecter	ENG20-01344236-v	recueillir, réunir, collecter	get or bring together	0.88
examiner	ENG20-00786734-v	concerner, présenter, examiner, envisager	think about carefully; weigh	0.87
examiner	ENG20-02091359-v		to look at critically or searchingly, or in minute detail	0.87
examiner	ENG20-02104471-v	concerner, présenter, examiner, envisager	give careful consideration to	0.87
examiner	ENG20-02069480-v	examiner	observe, check out, and look over carefully or inspect	0.87
examiner	ENG20-00623929-v	analyser	consider in detail and subject to an analysis [...]	0.87
nettoyer	ENG20-00034523-v	nettoyer	clean one's body or parts thereof, as by washing	0.86
nettoyer	ENG20-00174453-v	nettoyer	remove unwanted substances from	0.86
nettoyer	ENG20-00039521-v		make oneself clean, presentable or neat	0.86
nettoyer	ENG20-01490091-v	nettoyer	remove while making clean	0.86
nettoyer	ENG20-02661729-v	nettoyer	be cleanable	0.86
nettoyer	ENG20-01490246-v	nettoyer, purifier	make clean by removing dirt, filth, or unwanted substances from	0.86
voter	ENG20-02388587-v	voter	express one's choice or preference by vote	0.85
propager	ENG20-02001681-v		move outward	0.84
propager	ENG20-00936422-v	propager, parsemer, diffuser	cause to become widely known	0.84
propager	ENG20-01338470-v	propager	distribute or disperse widely	0.84
tester	ENG20-02456388-v	tester, soumettre, charger	put to the test, as for its quality, or give experimental use to	0.83

Table 3: Top 20 candidats as ranked by the method based on semantic similarity. 17 of them are corrects, while the three erroneous canidates are coming from a false lemmatization of reflexive verbs: *se nettoyer* would be correctly positioned in the given synsets with a reflexive lemma and *se propager* is semantically very close to the suggested synset. Erroneous candidates are italicized.

a heuristic one: if only one empty synset existed among the synsets proposed for a verb, the verb was added to this synset. This method allowed us to fill 377 synsets (among which 45 were filled with more than one verb), triggered by 530 nominalization candidates already present in WOLF.

4.3. Semantic Similarity

The nominalization candidates for which this heuristic approach was not applicable were the ones that had more than one empty verbal synsets. These candidates were integrated into WOLF by a disambiguation method based on distributional similarity. Given a candidate {V,S} the adequacy of verb V for filling synset S was measured by combining two criteria.

The first criterion concerns the reliability of the nominalization itself, i.e. the distributional similarity between the noun and the verb: **nominweight**(v,s).[2] The second criterion concerns the semantic similarity between the verb and the synset.

For calculating the semantic similarity between a verb and a synset S, we represented each synset by a bag of words containing the words present in S and in the other synsets linked to S by at most three steps of distance in terms of hypernymy or hyponymy relations. The similarity of a verb to a synset (**semweight**(v,s)) was calculated by comparing the information acquired for the verb from the learning corpus and the bag of words characterizing the synset: the lowest figure among the similarity measures between the verb and each word from the bag of words representing the synset S was taken into account.

In order to make best use of these two metrics which we presume to correlate with the semantic closeness of the verb to the synset, we tried to find a way to combine them. We decided to use the MegaM classifier [3], a maximum entropy based algorithm, to associate a weight to the two features.

[2]Distributional similarities between nouns and nouns, as well as between nouns and verbs, were calculated on the same corpus using the data and the metrics described above.

[3]http://www.cs.utah.edu/~hal/megam/

Learning data was created from the {V,S} candidates which are known to be correct because the verb already figured in the proposed synset in WOLF. Negative examples were also constructed assuming, for the sake of the experiment, that if the synset is not empty but does not contain the verb proposed by our method, the proposition is incorrect. (This is obviously an approximation, since we were interested in finding out which of them were correct). The classifier was trained on this data with the features semweight, nomweight and their product; it assigned a score **weight**(v,s) to each of the 1 716 candidates.

5. Evaluation

We proceeded to a manual evaluation of the resulting {V,S} candidates, to verify whether the assignment of the verb to the synset was correct. Only candidates coming from correct nominalizations were considered at this phase, since the precision of nominalizations had already been evaluated (cf. section 3.5.).

First, we examined 63 randomly chosen synsets filled by the heuristic method. The errors due to incorrect nominalization candidates put aside, this method positions 95% of the candidates in the correct synset.

Second, we selected 93 candidates (coming from a correct nominalization), assigned by the disambiguation method. Due to the higher level of ambiguity in the input data (i.e. more than one empty synsets for each lemma), a lower precision can be expected for this task. As a baseline, we calculated the proportion of correct {V,S} assignments in the totality of {V,S} couples generated from derivational relations: this represented a precision of 63%. The efficiency of the disambiguation method can be perceived as its capacity to reduce this error rate by ranking candidates according to their weight score. The score proved to correlate with the precision of the assignment: 80% of the 25% highest ranked of candidates were correct, while the proportion drops to 69% when considering the top 50% highest-ranked candidates.

6. Conclusion

We presented an unsupervised method for extracting information about event nominalizations from monolingual corpora. The results of the method were exploited to improve the coverage of a lexical semantic resource, the WOLF. The proposed algorithm is based on distributional analysis: a set of verb-noun couples, with similar syntactic contexts and semantic role assignment, is extracted from our learning corpus. The distributional analysis is complemented by a morphological module and a filtering on the basis of an 'event indicator' measure.

The main advantage of our method over currently existing nominalization lexica is to be able to extract morphologically unrelated nominalizations, and to provide a set of contexts indicating the meaning in which the noun and verb are related and thus facilitate manual validation.

Two complementary methods were employed to integrate the nominalization candidates into WOLF. Prior to this, we used the derivational links present in the structure of WOLF to create {lemma, synset} candidates to be added to the resource. A heuristic approach was used to add new lemmata to verbal synsets. The method positioned a significant number of candidates into WOLF with high precision. A disambiguation method completes the procedure by dealing with candidates for which the heuristic method did not apply. This method assigns an adequacy score to each {lemma, synset} pair, and the synset in the highest ranked pair is filled by the verb. The manual evaluation showed that this measure correlates with the actual correctness of the candidate. Although the current results do not allow a fully automatic population of WOLF, the candidates resulting from our method considerably facilitate the manual extension of the lexical resource.

7. Acknowledgements

The work reported in this paper was accomplished within the project SCRIBO, funded by SYSTEM@TIC, and within the project ANR EDyLex (ANR-09-CORD-008).

8. References

Eneko Agirre and Aitor Soroa. 2007. Semeval-2007 task 02: Evaluating word sense induction and discrimination systems. In *Proceedings of the 4th International Workshop on Semantic Evaluations, ACL*, pages 7–12, Prague, Czech Republic.

Marianna Apidianaki and Tim Van de Cruys. 2011. A Quantitative Evaluation of Global Word Sense Induction. In *Proceedings of the 12th International Conference on Intelligent Text Processing and Computational Linguistics (CICLing)*, Tokyo, Japan.

Béatrice Arnulphy, Xavier Tannier, and Anne Vilnat. 2010. Les entités nommées événement et les verbes de cause-conséquence. In ATALA, editor, *Proceedings of the 17th TALN Conference*.

Cécile Fabre and Didier Bourigault. 2006. Extraction de relations sémantiques entre noms et verbes au-delà des liens morphologiques. In *Proceedings of the 13th TALN Conference*, pages 121–129, Leuven, Belgique.

Christiane Fellbaum, editor. 1998. *WordNet. An Electronic Lexical Database*. The MIT Press, Cambridge, MA ; London.

Zellig Harris. 1954. Distributional structure. *Word*, 10(23):146–162.

Nabil Hathout, Fiammetta Namer, and Georgette Dal. 2002. An Experimental Constructional Database: The MorTAL Project. In Paul Boucher, editor, *Many Morphologies*, pages 178–209. Cascadilla, Somerville, Mass.

Marti A. Hearst. 1992. Automatic acquisition of hyponyms from large text corpora. In *Proceedings of the 14th International Conference on Computational Linguistics (COLING-92)*, pages 539–545, Nantes, France.

Mathieu Lafourcade and Alain Joubert. 2008. Détermination des sens d'usage dans un réseau lexical construit grâce à un jeu en ligne. In *Actes de TALN'08 15e Conférence Annuelle sur le Traitement Automatique des Langues Naturelles (TALN-2008)*, Avignon, France.

Maria Lapata. 2002. The disambiguation of nominalisations. *Computational Linguistics*, 28(3):357–388.

Cédric Messiant, Kata Gábor, and Thierry Poibeau. 2010. Acquisition de connaissances lexicales à partir de corpus : la sous-catégorisation verbale en français. *Traitement Automatique des Langues*, 51(1).

Claire Mouton and Gaël de Chalendar. 2010. JAWS: Just Another WordNet Subset. In *Proceedings of the 10th TALN Conference*, Montreal, Canada.

Sebastian Padó, Marco Pennacchiotti, and Caroline Sporleder. 2008. Semantic role assignment for event nominalisations by leveraging verbal data. In *Proceedings of the 22nd International Conference on Computational Linguistics - Volume 1*, COLING '08, pages 665–672, Stroudsburg, PA, USA. Association for Computational Linguistics.

Patrick Pantel and Dekang Lin. 2002. Discovering word senses from text. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 613–619, Edmonton, Alberta, Canada.

Benoît Sagot and Darja Fišer. 2008. Combining Multiple Resources to Build Reliable Wordnets. In *Proceedings of the 11th International Conference on Text, Speech and Dialogue (TSD 2008)*, Brno, Czech Republic.

Benoît Sagot, Karen Fort, and Fabienne Venant. 2009. Extension et couplage de ressources syntaxiques et sémantiques sur les adverbes. *Linguisticæ Investigationes*, 32(2):305–315.

Ludovic Tanguy and Nabil Hathout. 2002. Webaffix : un outil d'acquisition morphologique dérivationnelle à partir du Web. In *Proceedings of the 9th TALN Conference*, pages 245–255, Nancy France.

Jean Véronis. 2004. Hyperlex: lexical cartography for information retrieval. *Computer Speech & Language*, 18(3):223–252.

Éric Villemonte de la Clergerie. 2010. Convertir des dérivations TAG en dépendances. In *Actes de TALN'10 17e Conférence Annuelle sur le Traitement Automatique des Langues Naturelles (TALN-2010)*, Montreal, Canada.

Anne Vilnat, Patrick Paroubek, Éric Villemonte De La Clergerie, Gil Francopoulo, and Marie-Laure Guénot. 2010. PASSAGE Syntactic Representation: a Minimal Common Ground for Evaluation. In *Seventh international conference on Language Resources and Evaluation - LREC 2010*, Valletta, Malta.

Piek Vossen, editor. 1999. *EuroWordNet : a multilingual database with lexical semantic networks for European Languages*. Kluwer, Dordrecht.