Boosting Variant Recognition with Light Semantics

Cécile Fabre
ERSS / Dépt de Sciences du Langage
Univ. Toulouse-Le Mirail
5 allées A. Machado
31058 Toulouse Cedex, France
cfabre@univ-tlse2.fr

Christian Jacquemin
CNRS-LIMSI
BP 133
91403 ORSAY Cedex
France
jacquemin@limsi.fr

Abstract
A reasonably simple, domain-independent, large-scale approach of lexical semantics to paraphrase recognition is presented in this paper. It relies on the enrichment of morpho-syntactic rules and the addition of four boolean syntactic-semantic features to a set of 1,023 words. It results in a significant enhancement of precision of 30% with a slight decrease in recall of 10%.

1 Overview
The recognition of paraphrases and variants is an important issue in several areas of information retrieval and text understanding. Merging paraphrastic sentences improves summarization by avoiding redundancy (Barzilay et al., 1999). Term variant conflation enhances recall in information retrieval by pointing at documents that contain linguistic variants of query terms (Arampatzis et al., 1998).

In (Jacquemin and Tzoukermann, 1999), a technique is proposed for the conflation of morpho-syntactic variants that relies solely on morphological and low-level syntactic features (part-of-speech category, number agreement, morphological relationships, and phrase structure). An analysis of these results shows the limitation of this approach: correct and incorrect variants cannot be separated satisfactorily on a purely morpho-syntactic basis. Some additional lexical semantics must be taken into consideration.

In this study we propose a reasonably simple, domain-independent, large-scale approach of lexical semantics to noun-to-verb variant recognition. It relies on the mere addition of two boolean syntactic features to 449 verbs and two boolean morpho-semantic features to 574 nouns. It results in a significant enhancement of precision of 30% with a slight decrease in recall of 10%. This new approach to semantics—human-based, efficient, involving simple linguistic features—convincingly illustrates the positive role of linguistic knowledge in information processing. It confirms that verbs and their semantics play a significant role in document analysis (Klavans and Kan, 1998).

2 Morpho-syntactic Approach to Nomino-verbal Variation
In order to illustrate the contribution of semantics to the detection of paraphrastic structures, we focus on a specific type of variation: the verbal variants of Noun-Preposition-Noun terms or compounds in French. For example, les contraintes restituées dans les coques sont analysées (the residual constraints in the shells are analyzed) is such a verbal variant of analyse de contrainte (constraint analysis).

As a baseline for the extraction of these variants, we use a set of five morpho-syntactic transformations for Noun-Preposition-Noun terms reported in (Jacquemin and Tzoukermann, 1999) (see Table 1).\(^1\) We use the notation \(\mathcal{M}(N_i)_V\) for the morphological link between the initial term and the transformed structure. It represents any verb in the same morphological family as \(N_i\). For instance, in English, and according to the CELEX database, \(\mathcal{M}(\text{analysis})_V = \{\text{to analyze}, \text{to psychoanalyze}\}\).

Given a \(N_1 P_2 N_3\) structure, these transformations are obtained through corpus-based tuning

\(^1\)The following symbols are used for syntactic categories: N(noun), A(adjective), Av(adverb), V(verb), C(coordinating conjunction), P(preposition), and D(determiner). In the regular expressions, ? denotes optional and | disjunction. Morphologically related words are underlined.
Table 1: Morpho-syntactic (MS) Variants of N₁ P₂ N₃ Terms

| NheadToV:  | \( \mathcal{M}(N_1)_V (Av^J (P^J D) | P D^J) A^J) N_3 \)          |
|------------|----------------------------------------------------------|
| stabilité de prix (price stabilization) | -> stabiliser leurs prix (stabilize their prices) |

<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>abattage d'arbre (tree cutting)</td>
<td>-&gt; arbres ont été abattus (trees have been cut down)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>NmodifToV1:</th>
<th>N₁ ((A^J A (CAv^J A)^J V^V P) \mathcal{M}(N_3)_V</th>
</tr>
</thead>
<tbody>
<tr>
<td>méthode d'évaluation (method of evaluation)</td>
<td>-&gt; méthode pour évaluer (method for evaluating)</td>
</tr>
</tbody>
</table>

| NmodifToV2:  | N₁ (A^J (V | (P D^J (Av^J A)^J N^J) (Av^J A)^J Av^J) \mathcal{M}(N_3)_V          |
|--------------|--------------------------------------------------------------------------------------------------|
| zone de déstabilisation (region of destabilization) | -> zone déstabilisée (destabilized region) |

| NmodifToVRev:  | \( \mathcal{M}(N_3)_V (Av^J (P^J D) | (P D^J)A^J) N_1 \)                             |
|----------------|----------------------------------------------------------|
| température de chauffage (temperature for heating) | -> chauffés à haute température (heated at high temperatures) |

and correspond basically to four configurations:

1. either N₁ or N₃ (respectively head and modifier of the initial term) is transformed into a morphologically related verb V,
2. the order of the two content words is retained or reversed,
3. the dependency relation between the two initial nouns is preserved.

For instance, rule NheadToVRev corresponds to transformations in which the head noun is morphologically related to the verb and the order of the two words is reversed; rule NmodifToV (modifier transformed, order retained) has been divided into two subrules: the first one - NmodifToV1 - requires the insertion of a preposition just before the verbal form.

3 The Limits of the Morpho-syntactic Approach

In the first step of this work, we expected the precision of variant recognition to be controlled in two ways: firstly, by searching for multi-term variants in which the two content words of the initial term are found, directly or via morphological transformation. Secondly, by defining morpho-syntactic patterns of variation in terms of part-of-speech strings that are allowed to come in between these two content words.

Yet, the sequences found on such a morpho-syntactic basis prove to be of varying quality regarding their ability to provide paraphrases of the initial term. Consider for instance some of the variants detected for the term comparaison de résultat (comparison of results), in which only the first two sequences are good variants:

- compare les résultats (compare the results) (rule NheadToV, pattern \( \mathcal{M}(N_1)_V DN_3 \))
- résultats expérimentaux sont comparés (experimental results are compared) (rule NheadToVRev, pattern N₃AV\( \mathcal{M}(N_1)_V \))
- comparés aux résultats (compared to the results) (rule NheadToV, pattern \( \mathcal{M}(N_1)_V PN_3 \))

résultat d'une comparaison (resulted from a comparison) (rule NModifToVRev, pattern \( \mathcal{M}(N_3)_V PDN_1 \))

Such examples show that morpho-syntactic patterns are too coarse-grained to ensure that the dependency relation between the two pivots (results is the object of the predicate comparison) is maintained. When trying to define linguistic criteria to evaluate such variants, it appears that the frontier between good and bad variants lies between those that preserve the argument relation between the two content words and those that disrupt it. This means that, in the verbal variant, the argument relation between the verb and the noun must be the same as the relation between the deverbal noun and the other noun in the nominal term.

None of the five rules ensures that the subcategorization frame is preserved. For instance, if we consider the rule NModifToVRev, we find se-
quences that obey this constraint and sequences that violate it:

\[
\text{critère d'évaluation (evaluation criterion) } \rightarrow \text{évalué selon les critères (evaluated according to the criteria)}
\]

\[
\text{système d'évaluation (evaluation system) } \rightarrow \text{évalué le système (evaluated the system)}
\]

In the second case, the transformation is unacceptable because the instrumental relation expressed in the nominal term becomes an object relation in the verbal sequence. Even when word order is preserved, the relation between the pivots can be totally different in the term and its transformation, as in: \text{contrôle d'installation} (installation control) and \text{contrôle centralisé installé} (installed centralized control) (rule \text{NModifToV2}).

Our aim was to formulate additional constraints in order to control argument structure preservation. We thus had to cope with problem of handling nominal phrases (NP) in which one of the elements is morphologically linked to a verb. In French, as in English, the semantics of these nominal phrases is an issue for linguistic description: the two nouns can be linked by the whole range of argument-predicate relations, and very few linguistic elements can be used to decide what relation is expressed. Here is a brief list of the configurations that are likely to appear in such NPs:

- the second noun is the object of the first one: \text{comparaison de résultat} (comparison of result)
- the second noun is the subject of the first one: \text{augmentation de l'intensité} (increase in intensity)
- the second noun is an adjunct: \text{traitement à la chaleur} (treating with heat)
- the first noun is an adjunct: \text{taux d'augmentation} (increase rate)

Our aim was to find a way to use surface linguistic knowledge, as required in such an area of NLP, to deal with the interpretation of these phrases.

4 Light Semantics for Nomino-verbal Variations

Our approach consisted of two steps: firstly, defining semantic clues for accepting or discarding variants and, secondly, defining new variation patterns based on these features.

4.1 Filtering Criteria

First, using linguistic results on the semantics of French NPs (Fabre, 1996; Bartning, 1990), we identified predicate-argument configurations that cannot be matched by a given pattern (‘reject’ heuristics in the sense of (Lapata, 1999)). For example, when rule \text{NModifToVRev} applies, \text{N}_1 \text{ de } \text{N}_3 \text{ terms cannot be paraphrased by verbal sequences in which } \text{N}_1 \text{ is the object of the verb, as in: } \text{expérience d'utilisation} \rightarrow \text{utilisaît une expérience} \text{ (used an experiment). In such a configuration, only non-thematic arguments (adjuncts) of the deverbal noun may be found inside the NP.}

Similarly, when rule \text{NheadToVRev} applies, \text{N}_1 \text{ de } \text{N}_3 \text{ terms cannot be paraphrased by verbal sequences in which } \text{N}_1 \text{ is the subject of a transitive verb, as in: } \text{utilisation de l'expérience} \rightarrow \text{expérience utilisant} \text{ (experiment using).}

This configuration provides variants only when the verb is intransitive or ergative: ergative verbs allow for alternations of the form: \text{NP V (la densité augmentée) / one V NP (on augmente la densité).}

In this case, the following transformation is correct: \text{augmentation de densité (density increase)} / \text{densité augmentée (density increases).}

4.2 Enriched Metarules

Once it has been established which transformations should be rejected, we searched for surface linguistic clues that could help us to filter out these undesirable variants. It led us to the redefinition of the metarules, in two ways: putting additional constraints on the part-of-speech strings that can intervene between the two pivots, and defining new features to add linguistic control upon the application of the rules. These features are: the prepositional form, the morphological type of the noun, the transitivity of the verb, and the voice (active versus passive).

Here are two examples for the redefinition of the metarules (further details and examples are given in table 3):

\text{rule NModifToVRev} In this case, the metarule is transformed into a single
refined rule, in which the combination of parts of speech is more restricted: a preposition is required to eliminate object relations from the verbal phrase. In addition, the morphologically complex noun must be a processive deverb. Transformations such as expérience d’utilisation \( \rightarrow \) utilisait une expérience are filtered out.

**rule NheadToVRev** Here, the initial metarule is refined into three enriched rules, mainly by means of lexical constraints on the verb form. Only \( N_1 \) \( P_2 \) \( N_3 \) terms where \( P_2 = \text{de} \) are treated. If the verb is transitive, then the verb form must be a past participle (rule NheadToVRev-ActSimp (resp. NheadToVRev-ActComp) for simple (resp. complex) verb forms). Transformations such as utilisation de l’expérience \( \rightarrow \) expérience utilisant are filtered out.

The refinement of the metarules introduced four linguistic features which had to be encoded in the lexicon (see Table 2), namely:

- the morphological nature of the noun: the noun is either non deverb or deverb. In the latter case, it may correspond to an agent deverb, which refers to the agent of the verb, e.g. utilisateur (user), or to a processive deverb, which refers to the action denoted by the verb, e.g. utilisation (use).

- the transitivity of the verb: intransitive and ergative verbs are marked in the lexicon.

This annotation task is not time-consuming (about 3 hours for 1,023 words) and could be partly automated: characteristic endings could help to detect processive and agent deverbals. In addition, intransitive and ergative verbs form a small set of the verbal lexicon (8% of the verbs) which is likely to be partly domain-independent.

5 Experiments and Evaluations

In this section, we evaluate the variations produced from the two preceding sets of metarules: initial morpho-syntactic variations (henceforth MS) and new variations enriched through light semantics (henceforth MS+S).

The variants are obtained from a 13.2 million-word corpus composed of scientific abstracts in the agricultural domain (in French) and a set of 11,452 terms. The corpus is analyzed through SYLEX, a shallow parser that builds limited phrase structures and associates each word with an unambiguous syntactic category and a lemma. Terms are acquired from the output of the SYLEX parser. Only \([N][PN]\) structures are selected and only terms that occur at least three times in the corpus are retained.

The numbers of variants extracted through MS and MS+S are reported in Table 4. They are arranged in such a way that corresponding variations are aligned horizontally. For instance, each of the three MS+S variations NheadToV-Comp, NheadToV-Simp or NheadToV-Prep is a refinement of the MS variation NheadToV. In other words, the set of variants extracted by these three rich metarules is included into the set of variants extracted by the poor metarule. These two sets are not equal since the rich metarules are made more selective than the original metarule from which they are derived.

In addition to the output of rich and poor metarules, Table 4 shows, in the third column, the number of co-occurrences associated with these metarules. Co-occurrences are the least selective filters associated with morphosyntactic variants; they are expected to extract all the possible correct nominal- verbal variations (recall value 1.0). Given a \( N_1 \) \( P_2 \) \( N_3 \) term, these co-occurrences correspond to a configuration in which \( N_1 \) co-occurs with a verb that is morphologically related to \( N_3 \) or \( N_3 \) co-occurs with a verb related to \( N_1 \). Co-occurrences are extracted from a 11-word window (9 intervening words). These co-occurrences are used to evaluate the recall values of the filtering metarules.

---

3We are grateful to Xavier Polanco, Jean Royauté and Laurent Schmitt (INIST-CNRS) for providing us with this scientific corpus.
Table 2: Semantically Enriched Lexicon.

<table>
<thead>
<tr>
<th>Word</th>
<th>Processive Deverbal</th>
<th>Agent Deverbal</th>
<th>Intransitive</th>
<th>Ergative</th>
</tr>
</thead>
<tbody>
<tr>
<td>abaissier</td>
<td>−D</td>
<td>−A</td>
<td>−I</td>
<td>−E</td>
</tr>
<tr>
<td>abaissement</td>
<td>+D</td>
<td>−A</td>
<td>−I</td>
<td>−E</td>
</tr>
<tr>
<td>absorber</td>
<td>−D</td>
<td>−A</td>
<td>−I</td>
<td>−E</td>
</tr>
<tr>
<td>absorbeur</td>
<td>+D</td>
<td>+A</td>
<td>−I</td>
<td>−E</td>
</tr>
<tr>
<td>accorder</td>
<td>−D</td>
<td>−A</td>
<td>−I</td>
<td>−E</td>
</tr>
<tr>
<td>accord</td>
<td>+D</td>
<td>−A</td>
<td>−I</td>
<td>−E</td>
</tr>
<tr>
<td>accumulate</td>
<td>−D</td>
<td>−A</td>
<td>−I</td>
<td>−E</td>
</tr>
<tr>
<td>accumulation</td>
<td>+D</td>
<td>−A</td>
<td>−I</td>
<td>−E</td>
</tr>
<tr>
<td>accélérer</td>
<td>−D</td>
<td>−A</td>
<td>−I</td>
<td>+E</td>
</tr>
</tbody>
</table>

Table 3: Semantically Enriched Morpho-syntactic (MS+S) Variants of N1P2N3 Terms

**NheadToV-Comp:** \( \text{avoir} \, Av^2 \, \mathcal{M}(N_1)_{\text{V}} \, Av^2 \, DA^2 \, N_3 \)
\[
\{ \langle N_1 \, \text{dev} \rangle = \text{processive} \land P_2 = \text{de} \land \langle \mathcal{M}(N_1)_{\text{V}} \, \text{tense} \rangle = \text{past participle} \}
\]
comparaison de résultats (comparison of results)

**NheadToV-Simp:** \( \mathcal{M}(N_1)_{\text{V}} \, Av^2 \, DA^2 \, N_3 \)
\[
\{ \langle N_1 \, \text{dev} \rangle = \text{processive} \land P_2 = \text{de} \land \langle \mathcal{M}(N_1)_{\text{V}} \, \text{tense} \rangle \neq \text{past participle} \}
\]
évaluation de risques (evaluation of risks) \( \rightarrow \) évaluer les risques (to evaluate risks)

**NheadtoV-Prep:** \( \mathcal{M}(N_1)_{\text{V}} \, Av^2 \, P_2 \, DA^2 \, N_3 \)
\[
\{ \langle N_1 \, \text{dev} \rangle = \text{processive} \}
\]
exposition à la lumière (exposure to light) \( \rightarrow \) exposées à la lumière (exposed to light)

**NheadtoVRev-Pass:** \( N_3 (A^7 \, (PA^7 \, N \, (A \, (CA)^7)^7)^7 \, (CD^7 \, Av^2 \, A^7 \, NA^7)^7 \, V^7 \, \text{être}^7 \, Av^2)^7 \, \mathcal{M}(N_1)_{\text{V}} \)
\[
\{ \langle N_3 \, \text{agreement} \rangle = \langle \mathcal{M}(N_1)_{\text{V}} \, \text{agreement} \rangle \land P_2 = \text{de} \land \langle N_1 \, \text{dev} \rangle = \text{processive} \land \\
\langle \mathcal{M}(N_1)_{\text{V}} \, \text{tense} \rangle = \text{past participle} \land \langle \mathcal{M}(N_1)_{\text{V}} \, \text{valence} \rangle = \text{transitive} \}
\]
répartition de charge (weight distribution) \( \rightarrow \) charge également répartie (equally distributed weight)

**NheadtoVRev-ActSimp:** \( N_3 (A^7 \, (PA^7 \, N \, (A \, (CA)^7)^7)^7 \, (CD^7 \, Av^2 \, A^7 \, NA^7)^7 \, V^7 \, \text{être}^2 \, Av^2)^7 \, \mathcal{M}(N_1)_{\text{V}} \)
\[
\{ P_2 = \text{de} \land \langle N_1 \, \text{dev} \rangle = \text{processive} \land \langle \mathcal{M}(N_1)_{\text{V}} \, \text{valence} \rangle \neq \text{past participle} \land \\
\langle \mathcal{M}(N_1)_{\text{V}} \, \text{valence} \rangle = \text{ergative/intransitive} \}
\]
chute de température (drop in temperature) \( \rightarrow \) température chute (temperature drops)

**NheadtoVRev-ActComp:** \( N_3 (A^7 \, (PA^7 \, N \, (A \, (CA)^7)^7)^7 \, (CD^7 \, Av^2 \, A^7 \, NA^7)^7 \, \text{avoir}^2 \, Av^2)^7 \, \mathcal{M}(N_1)_{\text{V}} \)
\[
\{ P_2 = \text{de} \land \langle N_1 \, \text{dev} \rangle = \text{processive} \land \langle \mathcal{M}(N_1)_{\text{V}} \, \text{tense} \rangle = \text{past participle} \land \\
\langle \mathcal{M}(N_1)_{\text{V}} \, \text{valence} \rangle = \text{ergative/intransitive} \}
\]
fermentation de jus (juice fermentation) \( \rightarrow \) jus de raisins fermentés (fermented grape juice)

**Precision and Recall**

In order to calculate the precision and recall of the rich and poor metarules and to estimate the gains of semantic enrichment, a set of 1,000 co-occurrences has been randomly chosen among the 159,898 co-occurrences retrieved by the system. They have been divided into three sets: \( S_1 \) (500 co-occurrences) and \( S_2 \) and \( S_3 \) (250 co-occurrences). \( S_1 \) has been evaluated independently by the two judges (i.e. the two authors)
in order to test the level of agreement and $S_1$ and $S_2$ have been evaluated separately by only one judge each. Each cooccurrence has been marked as positive (a correct variation), negative (an incorrect variation) or inevaluable. Inevaluable cases correspond either to tagging errors or to incorrect terms such as *coque de forme* (shell of shape) which is an incomplete term structure because it should be followed by an adjective such as *coque de forme ovale* (oval-shaped shell). Only the cases of agreement between the two judges are used for the computation of recall and precision values.

The addition of semantics results in an increase of precision of 0.29: from 0.499 for MS to 0.789 for MS+S. The corresponding decrease of recall is much smaller: 0.11 from 0.696 for MS to 0.586 for MS+S. Precision and recall can be combined into a single measure such as the effectiveness measure $E_\alpha$ given by Formula (1) in which $\alpha$ is a parameter ($0 \leq \alpha \leq 1$) (van Rijsbergen, 1975):

$$E_\alpha = 1 - \frac{1}{\alpha \left( \frac{1}{n} \right) + (1 - \alpha) \left( \frac{1}{n} \right)}$$  \hspace{1cm} (1)

$E_\alpha$ varies from 0 to 1.0. Low values of $E_\alpha$ correspond to combined high recall and high precision. If we use $\alpha = \frac{1}{2}$ in order to assign an equal importance to precision and recall, the $E_{\frac{1}{2}}$ values are 0.419 for MS and 0.327 for MS+S. They indicate that the addition of semantics has significantly improved the quality of variant extraction. Detailed values of recall and precision are shown in Table 5.

**Agreement on Judgment**

Agreement on a classification task can be measured through the kappa coefficient ($K$). It evaluates the pairwise agreement among a set of coders making category judgment, correcting for expected chance agreement (Carletta, 1996). In our case the results of the ternary classification task are given by Table 6. The simple kappa coefficient is

$$K = \frac{P_0 - P_e}{1 - P_e}$$  \hspace{1cm} (2)

in which $P_0 = \sum_i \frac{n_i}{n}$ and $P_e = \sum_i (\frac{n_i}{n} \cdot \frac{n_i}{n})$ (Cohen, 1960). $P_0$ is the proportion of times the coders agree and $P_e$ is the proportion of times we would expect them to agree by chance. The value of the kappa coefficient is 0.91 indicating a good reliability of the evaluation performed by the two independent judges.

**6 Conclusion**

On a linguistic point of view, this experiment demonstrates that NLP applications can provide new issues for the description of linguis-
Table 5: Precision and recall in variant extraction for MS and MS+S variations

<table>
<thead>
<tr>
<th>P_{MS}</th>
<th>P_{MS+S}</th>
<th>R_{MS}</th>
<th>R_{MS+S}</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.438 NheadToV</td>
<td>0.938 NheadToV-Comp</td>
<td>0.875</td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.565 NheadToV-Simp</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.735 NheadToVRev</td>
<td>0.902 NheadtoVRev-Pass</td>
<td>1.000</td>
<td></td>
</tr>
<tr>
<td></td>
<td>— NheadtoVRev-ActSimp</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.111 NmodifyToV1</td>
<td>0.308 NmodifyToV1-Prp</td>
<td>0.357</td>
<td></td>
</tr>
<tr>
<td></td>
<td>1.000 NmodifyToV2-Inf1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.769 NmodifyToV2</td>
<td>— NmodifyToV2-Inf2</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>— NmodifyToV2-Inf3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.448 NmodifyToVRev</td>
<td>0.000 NmodifyToVRev-Prep</td>
<td>0.765</td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.765</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

0.499 0.789 0.696 0.586

Table 6: Frequencies of pairwise judgments for the ternary classification of nominal-verb variations (\* = inevaluable, + = correct, – = incorrect).

<table>
<thead>
<tr>
<th>n_{ij}</th>
<th>*</th>
<th>+</th>
<th>–</th>
<th>n_{i}</th>
</tr>
</thead>
<tbody>
<tr>
<td>*</td>
<td>120</td>
<td>9</td>
<td>1</td>
<td>130</td>
</tr>
<tr>
<td>+</td>
<td>1</td>
<td>184</td>
<td>6</td>
<td>191</td>
</tr>
<tr>
<td>–</td>
<td>4</td>
<td>10</td>
<td>165</td>
<td>179</td>
</tr>
<tr>
<td>n_{ij}</td>
<td>125</td>
<td>203</td>
<td>172</td>
<td>500</td>
</tr>
</tbody>
</table>

References


