Dictionary-Ontology Cross-Enrichment Using TLFi and WOLF to enrich one another

Emmanuel Eckard, Lucie Barque, Alexis Nasr, Benoît Sagot

Dictionary-Ontology Cross-Enrichment
Using TLFi and WOLF to enrich one another

Emmanuel ECKARD[1] *Lucie BARQUE*[2] *Alexis NASR*[1] *Benoît SAGOT*[3]
(1) Laboratoire d'Informatique Fondamentale de Marseille, UMR 7279 - CNRS, Université Aix Marseille
(2) LDI, UMR 7187- CNRS, Université Paris 13, France
(3) Alpage, INRIA Paris-Rocquencourt & Université Paris 7, France
Emmanuel.Eckard@a3.epfl.ch, lucie.barque@univ-paris13.fr,
benoit.sagot@inria.fr, Alexis.Nasr@lif.univ-mrs.fr

ABSTRACT
It has been known since Ide and Veronis [6] that it is impossible to automatically extract an ontology structure from a dictionary, because that information is simply not present. We attempt to extract structure elements from a dictionary using clues taken from a formal ontology, and use these elements to match dictionary definitions to ontology synsets; this allows us to enrich the ontology with dictionary definitions, assign ontological structure to the dictionary, and disambiguate elements of definitions and synsets.

KEYWORDS: Dictionaries, ontologies, WordNet.

1 Introduction

It has been known since Ide and Veronis [6] that it is impossible to extract an ontology structure from a dictionary, because this information is simply not present in the dictionnary as it is in the ontology, not even implicitly. Human intuition that dictionary definitions contain an ontology-like structure stems from the world knowledge that we uncounsciously also take into consideration as context when we read them; since this world knowledge is not available to computers, automated extraction fails. For instance, one of the Wiktionary definitions for "lock" is "A segment of a canal or other waterway enclosed by gates, used for raising and lowering boats between levels". The term "canal" here is polysemous, defined either as "An artificial waterway, often connecting one body of water with another" or "A tubular channel within the body". A computer has no straightforward[1] way to tell which of the senses is relevant, while a human, linking "waterway", "boats" and "body of water" to a common semantic field through their experience of the world, will easily choose the boating sense over the anatomic one.

Since the closest thing to a world knowledge available to computers is precisely ontologies, it seems appealing to design an ontology-powered automated process to identify elements of ontological structure present in the dictionary. Clearly, the ontology information that we inject into the process should not excessively constrain it, lest we find that very information. Instead, the process should trigger a virtuous circle where clues from the ontology permit structuring the dictionary, which in turn enriches the ontology with dictionary information. Only under these conditions can the process be both practical and useful. The approach of nurturing

[1] Recognising a common semantic field for two segments of text that share few or no common words goes beyond mere co-occurrence count; it is feasible, but requires sophisticated strategies such as latent semantics, for instance, and is difficult on small samples.

Proceedings of the 3rd Workshop on Cognitive Aspects of the Lexicon (CogALex-III), pages 81–94,
COLING 2012, Mumbai, December 2012.

interpretation of dictionary definition with ontology information can be considered from two points of view: a minima, as relaxing the strong hypotheses of "dictionary information only", under which Ide and Veronis showed extraction to be impossible[2]; a maxima, as injecting information into the process to mimick Human understanding of definitions through world knowledge.

Resources containing world knowledge can provide their information in different formats: for instance, dictionaries provide a number of definitions for each given word, with a distinct definition for each sense, and possibly hierarchies of sub-meanings; ontologies also provide short definitions, but mostly provide a structured set of relationships between senses, such as hypernymy, meronymy, etc. Although a wealth of resources exists in computer-readable form, resources become scarcer when we consider languages other than English. For instance, the general ontologies WordNet and FrameNet in English are hand-built, quite complete and available under a Free software-like licence [3, 4]. In French, on the other hand, the most notable alternatives are Euro Wordnet, which is quite complete and hand-built but only available under a commercial licence [13], and WOLF, which is available under a Free licence but is computer-generated from WordNet and incompletely translated. WOLF particularly suffers from the difficulty to adequately identify and translate polysemous words [12].

Since it provides a great deal of information while leaving room for improvement, WOLF constitutes both a resource and a testing bed for new algorithms and heuristics. As a resource, we can use it to generate clues for our heuristic; as a testing bed, contribute improvements to it. In this work, we attempt to enrich WOLF with dictionary definitions taken from the TLFi (*Trésor de la Langue française informatisé*). Practically, this comes down to assigning a dictionary definition to ontology synset elements, or to match ontology synsets with precise senses of a word in the dictionary. To achieve this result, we will explore the ambiguous graph structure implicitly formed by TLFi definitions. The heuristic attempts to connect two words h and H through a hypernymy relation by recursively roaming the definitions of words contained in a definition, concentrating on a hypernym; when successful, it stores the list of elementary segments that connect h to H. For instance, WOLF predicts that *établissement* (establishment) is a hypernym of *académie* (academy); indeed, in TLFi, these words are connected through certain senses of *école* (school): we find

académie → école → établissement

The word *école* is contained in the definition of *académie* and its own definition in turn contains *établissement*. Hence, *académie* leads to *établissement* as predicted by the clue provided by WOLF. Each of the words visited by the heuristic yields a number of different senses, each with its own definition which is examined separately. Hence, the hierarchy actually detects

académie-6 → école-1 → établissement

After a successful connection attempt, the pairs of unique senses immediately connected to each other (like académie-6 → école-1) are recorded and a frequentation counter associated with the sense pair is incremented. The result of the process allows us to tell which sense of *école* is expressed in the definition of *académie* that we considered.

[2] Several studies proposed automatic or semi-automatic methods to develop lexical hierarchies from dictionary data, e.g. [2, 10].

2 Resources

2.1 TLFi

The *Trésor de la Langue française informatisé* (TLFi) [11] is the digital version of the *Trésor de la Langue française*, a large reference dictionary for French. The two main reasons why we have chosen the TLFi is that it is available in electronic form for research purpose and that most of its definitions belong to so-called *definitions by genus and differentiæ* allowing us to extract genus (or hypernym of the defined unit). The TLFi has also a wide coverage with around 270,000 definitions. This study is restricted to nouns, for which the TLFi provide 100,493 definitions describing the meaning(s) of 35,498 nominal entries.

The senses of a lexical entry in TLFi are subdivised into a hierarchy of senses and subsenses, each complete with a unique identification number and a definition; for instance, the word *bois* (wood) comprises the following senses[3]:

1.1.1 Ensemble d'arbres croissant sur un terrain d'étendue moyenne; ce terrain même.
2.1.1.1 Matière (racines, tronc, branches) qui constitue l'arbre (à l'exception du feuillage).

2.1.1 Identification of definitions genus

In the framework of the *Definiens* project, TLFi definitions of nouns were POS-tagged and processed to determine the *genus* of a given definition, that is, the noun or noun phrase that corresponds to the hypernym of the defined noun [1]. The Definiens heuristic relies on lexico-syntactic patterns that recognise nouns or noun phrases as possible genus candidates. More precisely, around fifty rules have been manually elaborated to identify geni in the TLFi definitions. Represented as finite-state transducers, the rules have been run on definitions previously labeled with part of speech tags by the NLP tool suite MACAON [8]. The rule presented in figure 1 identifies nominal definitions that begin with a common noun (nc for *nom commun* in French), followed by a preposition and then another noun (left hand side of the rule). The right hand side of the rule proposes two possible geni for this kind definition: the first noun or a more specific phrasal genus constituted by the three elements (noun, preposition, noun) detected in the left hand side of the rule. This rule matches for example the definition of JODHPURS presented below since it begins with a noun (*pantalon*) followed by a preposition (*de*) followed by a noun (*équitation*). The right hand side of the rule thus indicates two possible geni for this definition : *pantalon* (trousers) and *pantalon d'équitation* (horse riding trousers).

JODHPURS = Pantalon d'équitation importé des Indes par les officiers anglais, ajusté du genou à la cheville et qui se porte sans bottes (Horse riding trousers imported from India by English officers, tight from knee to ankle and that is worn without boots.) ⇒ **genus 1**: pantalon, **genus 2**: pantalon d'équitation

The rule presented in figure 1 also matches the definiton of BOIS-1.1.1 given above. Nevertheless, geni like "ensemble de N" have to be treated in a particular way. Thus, the rules can also include lexical elements, as illustrated below in figure 2: when a definition matches the

[3] "Wood" 1.1.1 Set of trees growing on a medium-sized area of land; said terrain.
2.1.1.1 Matter (roots, trunk, branches) that constitute a tree (except the foliage). This particular case is provided here to examplify definition numbering, without prejudice of further questions, like whether the "said terrain" metonymy should ideally be numbered separately. In the Princeton Wordnet, only the first half of this definition appears at all.

```
<rule>
  <lhs>
    <elt cat="nc"/>
    <elt cat="prep"/>
    <elt cat="nc"/>
  </lhs>
  <rhs>
    <genus><elt num="1"/></genus>
    <rhs>
       <genus>
         <elt num="1"/>
         <elt num="2"/>
         <elt num="3"/>
       </genus>
    </rhs>
  </rhs>
</rule>
```

Figure 1: Example of a syntactic genus extraction rule

sequence *ensemble de/d' + nc* (set of), the selected genus is not *ensemble* but the common noun that follows the preposition. The noun is moreover applied to the function "set of".

```
  <regle>
    <lhs>
      <elt lex="ensemble"/>
      <elt cat="prep"/>
      <elt cat="nc"/>
    </lhs>
    <rhs>
      <function><elt num="1"/></function>
      <genus><elt num="3"/></genus>
    </rhs>
  </regle>
```

Figure 2: Example of a lexico-syntactic genus extraction rule

When the rules propose multiple geni for a given word sense, as in the rule presented in figure 1 above, the genus that is selected is the most specific one, provided that this most specific genus is classifying (*i. e.* appears as a genus in at least another definition). In other words, the genus that is nor too specific nor too general is assumed to represent the most accurate genus. In the JODHPURS example, the genus *pantalon d'équitation* (horse riding trousers) is more specific than *pantalon* (trousers) but the processing of the whole corpus tells us that *pantalon d'équitation* appears only once, in the definition of JODHPURS, whereas *pantalon* appears in the definition of eighteen word senses (BLUE-JEAN, SAROUAL, ...) in the TLFi.

This automatic process, that consists in counting every possible geni of every definitions through the corpus, allows us to obtain the data described in table 1 below. As shown in

Nominal words	35,498
Nominal word senses	100,493
Distinct geni	17,204
Classifying geni :	13,924
Simple nouns	5,578
Phrasal nouns	8,346

Table 1: Geni extracted from the TLFi

this table, the 13,924 classifying geni are composed of 5,578 simple nouns (*e.g.* `conifère` (conifer), `formule` (formula), ...) and 8,346 phrasal nouns (*e.g.* `conifère de grande taille` (tall conifer), `courte formule` (short formula); ...). Let's recall that phrasal geni are very interesting in that they "naturally" disambiguate ambiguous forms (cf. `carte` *vs* `carte géographique` (map) and `carte à jouer` (playing card)). The 8,346 phrasal geni that have been yet detected are based on only 1,754 distinct nominal heads and more than 90 percent of them are included in the simple nouns set. The total number of words to disambiguate is therefore equal to 5,578 simple nouns plus 175 heads of complex nouns that are not already included in the simple geni set. Most of these geni are ambiguous, for an average of 4 senses per genus.

2.1.2 Adaptation of the sense hierarchy to limit ambiguity

The number of distinct senses can be high, and the differences between some of them can be quite subtle. For instance, the verb "to dive" distinguishes the senses "move briskly and rapidly downwards" and "being directed downwards". TLFi also records unusual or archaic senses of words: for instance, the term *fourchette* ("fork") lists the chess configuration, which might not be the first to spring to mind, as well as an archaic vernacular word for "bayonet".

In TLFi, the different senses of a word are organised in a hierarchy of sense numbers, such as "1", "2.3.1", etc. Senses with more decimals in their sense number are children of the parent sense, i.e. variants of the parent sense in a particular framework. Senses with 4 or more decimals in their sense number tend to be very specific senses, with long definitions. To avoid hyper-correction, we deem it adequate to trim this sense hierarchy, as the fine granularity achieved by human lexicographers is not a realistic goal for our automatic system [7]. We devise two simple schemes to this purpose: the "cut" scheme simply ignores all definitions whose sense number bears more than a given number of decimals; and the "merge" scheme deletes all definitions whose sense number bears more than a given number of decimals, but concatenates their definition to that of their direct parent. For instance, "cut 2" will retain senses "1", "2.1" and "3.1", but will eliminate senses "1.2.1", "2.1.1", "3.1.2.1", etc.; and "merge 1" will retain senses "1" and "2", and will eliminate sense "2.1" and "2.3.1.1" after concatenating their definition to the definition of sense "2". In cases where definitions are merged, their geni are stored in a vector, which allows us to take them into consideration one by one.

2.2 WOLF

WOLF (*WOrdNet Libre du Français*)[4] is a French-language ontology, automatically built from the Princeton WordNet (PWN) and various other resources [12]. Monosemous literals in the PWN 2.0 were translated using a bilingual French-English lexicon built from various multilingual resources. Polysemous PWN literals were handled by an alignment approach based on a multilingual parallel corpus. The synsets obtained from both approaches were then merged. The resulting resource, WOLF, preserves the hierarchy and structure of PWN 2.0 and contains the definitions and usage examples provided in PWN for each synset. Although new approaches are currently being used for increasing its coverage [5], WOLF is rather sparse, as information was not found for all PWN synsets by these automatic methods. Indeed, one of the difficulties in completing WOLF is to disambiguate the words contained in its synsets as to

[4] `http://alpage.inria.fr/~sagot/wolf.html`

allow a correct translation, since the level of polysemy is high.

In this work, we used the version 0.2.0 of the WOLF, in which 46,449 out of the 115,424 PWN 2.0 synsets are filled with at least one French literal. WOLF 0.2.0 contains 50,968 unique literals which take part in 86,235 (literal, synset) pairs, i.e., lexical entries (to be compared with the 145,627 such pairs in the PWN 2.0). Approximately half of these pairs are nouns, i.e., belong to nominal synsets.

Since the WOLF was created automatically using several distinct techniques, each (literal, synset) pair is associated with the set of techniques that suggested its creation, together with a technique-specific confidence measure. This information is used for filtering out (literal, synset) pairs with the lowest confidence scores. We defined two filters: a medium filter, which retain more candidates, and a strong filter, which retain only the most reliable candidates (cf. figures in the next section).

3 Using hypernymic paths for synset–definition matching

Our aim is to enrich WOLF and TLFi with one another, entailing that we need to assign specific definitions to given WOLF synsets. These synsets, or sets of synonyms, contain words that share a same meaning, but this meaning is yet not explicitly determined. As such, these words are ambiguous with respect to TLFi, and it is not straightforward to decide which of the TLFi definitions should be associated with them, if any. To solve this issue, we propose to use the two resources and compound them with a heuristic.

The heuristic attempts to connect two words with a hypernymy relation, and stores the senses through which the connection goes in case of success. At each step, a definition is associated with hypernym candidate words — typically the head of the genus of the TLFi definition, provided by a pre-processing of TLFi (see section 2.1.1 ; the senses of this word are explored recursively in a breadth-first search until the goal is reached.

The WOLF hypernymy hierarchy provides us with numerous hyponym–hypernym couples, including measures of confidence for these couples. The heuristic processes all these couples, storing the elementary steps that constitute successful hypernymy paths, and keeping track of their frequentation.

The nature of dictionary definitions — short bursts of text completely independent from one another — prevents us from using machine learning techniques. Instead, we take advantage of the graph structures that are explicitly expressed in the ontology, and to some extent implicitly in the dictionary. Given a word h and a hypernym H of h, we use a graph exploration technique to connect senses of h and H. We then record pairs to constitute the path between h and H. This provides us with a set of word sense pairs that TLFi puts in direct hypernymy relation. We can then use these pairs to populate WOLF: if two words w and W are deemed to have definitions d and D in direct hypernymy according to TLFi, and belong to synsets s and S in WOLF, these synsets also being in the hypernymy relation, then we can safely identify d to s and D to S.

To disambiguate the hyponyms of an (h, H) pair, we explore the graph by *hypernymic ascent*: we consider the different senses $h_1, \ldots, h_n$ that TLFi provides for h, and attempt to connect each of them to any of the senses of H. Inspired by [9], we propose a connection scheme whereby we jump from one word to a word of its definition, iteratively, until we reach the target H. In our implementation of the hypernymic ascent scheme, we select the *genus* of the

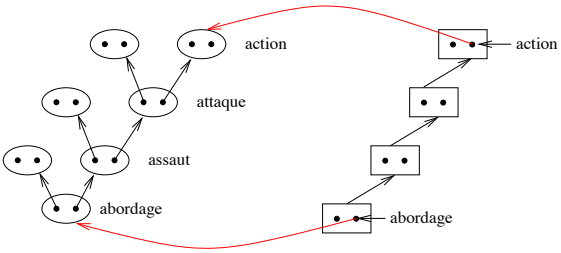

Figure 3: TLFi ambiguous structure (left) WOLF structure (right)

definition of a word (that can also be considered as its hypernym) to carry on the next iteration step, taking advantage of the preprocessing performed in the Definiens project [1].

This process is illustrated in figure 3. In the left hand part of the figure, we have represented the TLFi ambiguous structure. In this figure, the dots represent word senses while ellipses represent words. An ellipse that contains two dots therefore represent a polysemous word that has two possible different senses. An arrow linking a sense s (a dot) to an ellipse w (a word) indicates that the w is a hypernym of s. The problem, of course, is that we do not know which sense of w is actually the hypernym of s.

The right hand side of the figure represents the WOLF synset structure. Synsets are represented as rectangles while dots represent word senses. It must be noted that, in WOLF, word senses are not associated with definitions. In case of a polysemous word such that one of its senses is part of a synset, we do not actually know which sense it is. The arrows between rectangles represent the hypernymic relation.

In our example, we can extract from the WOLF subgraph that one sense of *abordage* has as a hypernym one sense of *action* although we do not know which sense of *abordage* nor which sense of *action* are linked by this relation. This is where the hypernymic ascent comes into play by looking, in the TLFi graph, for a path that links one sense of *abordage* with one sense of *action*.

The result of the hypernymic ascent is represented in figure 3. A path relating one sense of *abordage* to the word *action* has been discovered, it goes through a given sense of *assaut* (assault) as well as a given sense of *attaque* (attack). The number that labels the arcs between two senses corresponds to the number of paths that go through this arc.

Hypernymic ascent described can fail for several reasons. The main ones are described below:

1. Either the hypernym of the hyponym in an (h, H) pair extracted from the WOLF is absent from the TLFi. When used with the medium filter, a total number of $86,636$ (h, H) couples are extracted from the WOLF. For $49,908$ of them, both the hyponym and the hypernym are present in the TLFi. When the strong filter is used, $47,858$ couples are extracted out of which $24,443$ have both their hyponym and hypernym present in the TLFi.

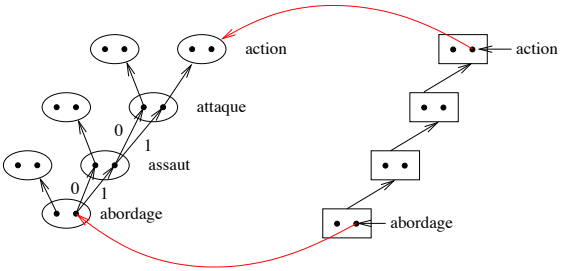

Figure 4: Result of the hypernymic ascent

2. Both h and H appear in the TLFi but no path was found that links them. This situation can have several causes :

   (a) Pre-processing errors. The preprocessing of the TLFi definition is made of several steps, each of which is error-prone. These steps are word segmentation of the definition, part of speech tagging and lemmatization.

   (b) Non standard definition. Although TLFi definitions generally follow a genus differentia schema some of them do not, some senses are defined, for example, by means of synonyms. In such cases the identification of the genus in the definition fails.

3. The (h, H) pair extracted from WOLF is incorrect. In such a case, a path can be found which contains at least one incorrect arc.

When the process actually succeeds, it can be the case that several paths are found that link h to H. A crude but quite effective way to deal with this situation is to select the shortest paths.

With the strong filter on WOLF hypernym couples (supposedly the most reliable set of (h, H) pairs given as clues), the success rate for connections is 21%; this falls to 18% with the medium filter (more details in table 2: for the strong and medium filters on WOLF (strong is the strictest and produces the most reliable couples), we give the number of words to explore, the number of senses yielded by the words, the number of successful connections, and the success rate of the connection attempts.).

	words	senses	success	success rate
medium	48,188	109,306	8,787	18.23%
strong	23,291	52,408	4,916	21.11%

Table 2: Connection attempts through hypernyms between two given words in hypernymic relationship.

The low success rate is ultimately neither a surprise, since a successful connection on one particular hyponym-hypernym pair is subject to many imponderables, nor a severe hindrance to our endeavour, since it is the accumulation of the elementary components yielded by the

successful connections that constitutes our result. Therefore, success rates around 20% are both well explained, and quite fit for our purpose.

It is worth noting that the scheme described above does not generalise as to disambiguate the hypernym in the (h, H) couple as well. This is ultimately due to a fundamental asymetry between the definitions of h and H: though the hyponym is often defined in terms that ultimately lead to the hypernym (either directly or through other definitions), the converse is not true since the hypernym H contains no information leading to the hyponym h. For example, a clue tells us that "snake" is the ultimate hypernym of "naja". The direct hypernym of "naja" is "cobra", which has several senses; only one of these senses has "snake" for hypernym, allowing us to discard cars and helicopters as candidate semantic fields. However, we have no way to determine which sense of "snake" is relevant (see figure 5).

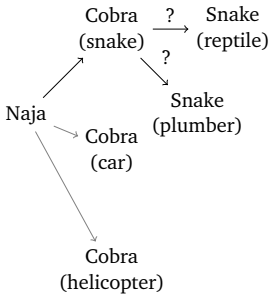

Figure 5: Ambiguous hypernym

One could attempt to reverse the hypernymic ascent into a hyponymic descent. However, this cannot be done simply by backtracking the path found during the hypernymic ascent, since the path is then completely determined. Hyponymic relationships linking TLFi dictionary definitions together should be available independently from the previously performed hypernymic ascent, for the hyponymic descent scheme to be viable. Unfortunately, this information is not present in TLFi. It is not possible to efficiently recreate this information by an exploratory preprocessing. For instance, envision a couple of direct hyponym-hypernym (w, W), such as one of the w_i is defined as being a kind of W; a pre-exploration of these relations would accurately detect that w_i is a hyponym of one of the $W_1, \ldots, W_n$, but it would have no direct way to tell the relevant W_i. In consequence, it is impossible to tell one W_i from another with this method, making hyponymic descent impractical with TLFi alone.

In summary, existence of quasi-hypernymic information in the form of geni of definitions featured in TLFi makes hypernymic ascent possible; absence of similar hyponymic data in definitions (like examples would partially provide) makes it impossible in practice to reverse the scheme.

4 Experiments

In order to measure the performances of the method, we ran it over two samples of 48,188 and 23,291 clues respectively (see table 2). After completion of the task, we randomly choose one hundred of the elementary hyponym-hypernym pairs and manually checked whether the

chosen senses for the hyponym and the hypernym are relevant *i.e.* we answer the question : "is H an appropriate hypernym of h?". The answer is yes for the "homme-10 / mâle-1" pair given below and no for the "verbe-4 / expression-1" pair :

- homme-10 = **Mâle** adulte de l'espèce humaine (adult male human) / mâle-1 = Individu appartenant au sexe qui possède le pouvoir de fécondation
- verbe-4 = **Expression** verbale de la pensée (à l'oral ou par écrit) (verbal expression of the thought) / expression-1 = Action d'extraire d'un corps le liquide qu'il contient (extraction of liquid from a substance)

We find a 45% accuracy in the tested sample. Given the average polysemy of 4.03 for Central Components in our sample, a random baseline will yield performance in the order of 25%; with our 45% accuracy, we are therefore significantly higher than the baseline.

The frequency of a segment (the number of times a segment appear in a successful path) did not correlate with the correctness of the segment. Instead, they tend to correlate with how high the segment is in the ontology, and thus to how general or abstract a segment is: many hypernymic paths tend to feature them as they climb towards the root of the ontology. Using them as an indicator for the correctness of a segment will need some kind of normalization with respect to the abstractness of the segment.

In order to get a better understanding of what happens during the hypernymic ascent, we present below a few examples of partially successful or failed paths.

academy – establishment, an unexpected and convoluted connection: We have seen a correct connection of "academy" to "establishment", through an adequate meaning of "school". Nevertheless, "academy" has no less than 15 meanings in TLFi. Notably, académie-18 is defined as "house of gaming or pleasure"[5]. This triggers a search through the heavily polysemic word *maison* (28 definitions) which eventually leads to "establishment" through

académie-18 → maison-41 → bâtiment-11 → grange-4 → établissement

Interestingly, all of the segments yielded by this search are actually valid. This is a good illustration of the fact that the connection of the terms of the WOLF clue is a mere pretext to the research of elementary segments: it does not matter much that the connection has taken a detour, as long as the elementary segments are valid – it can in fact yield more segments to enrich our collection.

baboon – animal, a connection through irrelevant definitions: WOLF predicts that *animal* (animal) is a hypernym of *babouin* (baboon); indeed, in TLFi, these words are connected through certain senses of *singe* (monkey) and *voyageur* (traveller): we find

babouin-1 → singe-24 → voyageur-14 → animal

By examining the definitions of these senses, we see there that the word *voyageur* ("traveller"), perhaps surprising at a first glance, is in fact taken in its acceptation of "moving animal"[6]; on

[5] *maison de jeu ou de plaisir*

[6] The definition for voyageur-14 gives "Animal roaming its natural habitat (air, sea, ground), particularly migratory birds" (*Animal se déplaçant dans son milieu naturel (air, mer, terre); en particulier, oiseau migrateur.*)

the other hand, the word *singe* ("monkey" or "ape") is taken in its unusual and little-known acceptation of "surnumerary passenger"[7], which is clearly not relevant in the context[8]. This case has successfully connected "baboon" to "animal", yet it yields two segments, babouin-1 → singe-24 and singe-24 → voyageur-14, that are both incorrect.

Similarly WOLF predicts that *adonis* (adonis) is a hyponym of *mâle* (male). One of the connections found is

adonis-7 → papillon-3 → personne-1 → individu-11 → homme-10 → male

Starting with the entomological sense of "adonis" (a *Lycaena* butterfly), we jump to "butterfly", but in the sense of "socialite"; from there, we follow a foreseeable path through "person", "individual", "man" and eventually "male". Here, the segment adonis-7 → papillon-3 is false, though the others are correct. Obviously, the overall path connecting the terms of the WOLF clue makes little sense to the Human eye, but this is less problematic than incorrect segments. The overall path is merely a pretext to the research of elementary segments. By contrast, another path found for the same clue is

adonis-8 → nom-30 → partie-31 → individu-11 → homme-10 → male

which makes more sense, but does not yield more correct elementary segments than the previous example.

steal mill – factory, a trivial connection: WOLF predicts that *aciérie* (steal mill) is a hyponym of *usine* (factory); indeed, in TLFi, the first and only definition of *aciérie* is "factory where steal is manufactured"[9], entailing that the connection is direct and trivial. Since the term *usine* has seven different definitions on TLFi, and since our heuristic leaves the ultimate hypernym ambiguous, it is impossible to select which sense of *usine* is relevant. Thus, in spite of a successful connection, this path yields no useful segment.

poster – worker, an erroneous WOLF clue: WOLF predicts that *affiche* (poster) is a hypernym of *ouvrier* (worker), a rather counter-intuitive pair; our heuristic manages to find a convoluted path that connects these two words, but it is clear that integrity of the semantic field has been lost en route. The connection path goes

affiche-8 → action-2 → mise-72 → investissement-1 → manœuvre-1 → ouvrier

The word *mise* is here taken as "stakes in a gamble", leading to "investment" taken in its economic sense; the sense of "investment" then switches to the military term for "surrounding an enemy", yielding the word *manœuvre* ("manoeuver"); *manœuvre* then switches to its meaning of "unqualified worker", eventually completing the connection. Yet, the segments mise-72 → investissement-1 and investissement-1 → manœuvre-1 are incorrect.

[7]The definition for singe-24 gives "Traveller installed on the upper floor out of a lack of space in the inside of a public car" (*Voyageur installé sur l'impériale faute de place à l'intérieur d'une voiture publique.*)

[8]One set of experiments considered ignoring archaic meanings, as well as all specialised meaning marked by a domain tag in TLFi, to aleviate ambiguity somewhat; this did not yield significant improvement in performance.

[9]*Usine où se fabrique l'acier*

5 Conclusions

We have described an exploration scheme of how the hypotheses of Ide and Veronis can be relaxed as to make it possible to automatically align a dictionary and an ontology. We use "clues" extracted from an ontology to search consistent paths in the dictionary linking a hyponym to a hypernym, recording the intermediary steps that form the overall path. We attempted this using WOLF and TLFi, taking advantage of the TLFi dataset that was made available to us.

This "hypernymic ascent" scheme yields a high rate of connection failures, which is not in itself a problem as these connections are a pretext to recording the elementary segments that form the connection. Nevertheless, this indicates that relying on Central Components to climb in the hypernymy chain is not very efficient in the context of a natural language dictionary. We could envision better performances using more rigidely formatted dictionaries and less naive approximations for the hypernym of a definition than merely using its Central Component.

Another issue is that words close to the root of the ontology tend to be very fundamental and highly polysemic. Therefore, a connection that passes through them is likely to have lost its semantic integrity. The yields semantically inconsistant segments, thereby generating noise.

In spite of the many difficulties that we encounter with the data and the naive nature of some elements of our system, we still managed to obtain a 45% accuracy on a randomly selected sample, significantly above the random baseline. This makes our system suitable as a weak classifier as it is, and leaves much room for improvement using more rigidely formatted and self-consistent data, better management of word inflexions, and refined selection of definition features beyond mere central components.

Acknowledgments

This work has been funded by the French Agence Nationale pour la Recherche, through the project EDYLEX (ANR-08-CORD-009).

References

[1] L. Barque, A. Nasr, and A. Polguère. From the definitions of the trésor de la langue française to a semantic database of the french language. In *European Association for Lexicography International Congress (EURALEX)*, Leeuwarden, Pays Bas, 2010.

[2] M. Chodorow, R. Byrd, and G. Heidorn. Extracting semantic hierarchies from a large on-line dictionary. In *ACL*, pages 299–304, 1985.

[3] C. Fellbaum. *WordNet: An Electronic Lexical Database*. MIT Press, Cambridge, Mass, 1998.

[4] C. Fillmore, C. Johnson, and M. Petruck. Background to Framenet. *International Journal of Lexicography*, 16:235–250, 2003.

[5] V. Hanoka and B. Sagot. Wordnet creation and extension made simple: A multilingual lexicon-based approach using wiki resources. In *Proc. of the 8th international conference on Language Resources and Evaluation (LREC)*, page 6, Istanbul, Turquie, 2012.

[6] N. Ide and J. Veronis. Extracting knowledge-bases from machine-readable dictionaries: Have we wasted our time? In *Proc KB&KB'93 Workshop*, 1993.

[7] N. Ide and Y. Wilks. Making sense about sense. In *Text, Speech and Language Technology*, volume 33, pages 47–73, 2006.

[8] A. Nasr, F. Béchet, J.-F. Rey, B. Favre, and J. Le Roux. Macaon: An nlp tool suite for processing word lattices. In *The 49th Annual Meeting of the Association for Computational Linguistics*, 2011.

[9] R. Navigli. Using cycles and quasi-cycles to disambiguate dictionary glosses. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, EACL '09, pages 594–602, Athens, Greece, 2009.

[10] R. Navigli and P. Velardi. From glossaries to ontologies: Extracting semantic structure from textual definitions. In *Ontology Learning and Population: Bridging the Gap between Text and Knowledge*, pages 71–87, 2008.

[11] J. M. Pierrel. Le Trésor de la Langue Française Informatisé : un dictionnaire de référence accessible à tous. *AMOPA*, (174):25–28, 2006.

[12] B. Sagot and D. Fišer. Automatic Extension of WOLF. In *GWC2012 - 6th International Global Wordnet Conference*, Matsue, Japon, January 2012. PHC PROTEUS 22718UC.

[13] P. Vossen. *EuroWordNet : a multilingual database with lexical semantic networks for European Languages*. Kluwer, Dordrecht, 1999.