Applying Cross-Lingual WSD to Wordnet Development

Marianna Apidianaki[1,2], **Benoît Sagot**[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
marianna.apidianaki@limsi.fr, benoit.sagot@inria.fr

Abstract

The automatic development of semantic resources constitutes an important challenge in the NLP community. The methods used generally exploit existing large-scale resources, such as Princeton WordNet, often combined with information extracted from multilingual resources and parallel corpora. In this paper, we show how Cross-Lingual Word Sense Disambiguation can be applied to wordnet development. We apply the proposed method to WOLF, a free wordnet for French still under construction, in order to fill synsets that did not contain any literal yet and increase its coverage.

Keywords: WordNet, Word Sense Disambiguation, Cross-Lingual Word Sense Induction

1. Introduction

The need for lexical and semantic knowledge in NLP applications has steered several initiatives for resource development in recent years. A general trend has been to develop multilingual semantic resources on the basis of Princeton WordNet (PWN) (Fellbaum, 1998) by preserving its structure and transferring its contents into new languages using various translation-based methods (Vossen, 1998; Pianta et al., 2002; Tufiş et al., 2004). This approach presents several advantages which explain its wide adoption. It mainly permits to avoid the time-consuming and expensive manual elaboration of the semantic hierarchy in new languages, and allows the alignment of the resulting wordnets between them and to PWN. Its weaknesses concern the bias imposed by PWN on the content and structure of the new wordnets, the manual work required for the transfer and the reliance on predefined lexico-semantic resources.

In an attempt to address these weaknesses, several wordnet development methods have been proposed that exploit information extracted from parallel corpora. These methods permit to automatically acquire semantic information from texts and thus circumvent the need for predefined resources, as well as the manual filling of new wordnets. Following this line of research, our aim is to show how Cross-Lingual Word Sense Disambiguation (CL-WSD) can be applied to wordnet development, for creating new resources or enriching existing ones. We illustrate this approach by applying the CL-WSD method of Apidianaki (2009) to the enrichment of an automatically-built wordnet for French, the WOLF (Sagot and Fišer, 2008).

The paper is organized as follows. Section 2 presents various, more or less supervised, wordnet development methods. Section 3 presents the semantic resource WOLF that we aim to enrich. In Section 4, we explain how unsupervised Word Sense Induction and Disambiguation can be applied to wordnet development. In Section 5, we present the results of a manual evaluation we carried out for estimating the quality of the new WOLF entries, before discussing some perspectives for future work.

2. Cross-lingual approaches to wordnet development

2.1. The expand model

Multilingual wordnet development has generally heavily relied on Princeton WordNet (PWN) (Fellbaum, 1998). Projects aiming the development of wordnets for languages other than English, such as EuroWordNet, BalkaNet and MultiWordNet (Vossen, 1998; Pianta et al., 2002; Tufiş et al., 2004), have widely adopted a translation-driven approach: the structure of PWN was preserved while its contents were imported in the newly built resources by applying various translation-based methods.

This approach, also called the *expand model*, permits to avoid the time-consuming and expensive manual elaboration of the semantic hierarchy in new languages. An additional advantage is that the resulting wordnet is automatically aligned to PWN, as well as to other wordnets built following the same principle. The resulting resources are thus interesting in a contrastive perspective and can be particularly useful in multilingual NLP tasks.

Despite its numerous advantages, the translation approach is also characterized by several weaknesses. First of all, the content and structure of the target wordnets are strongly biased by PWN, based on the assumption that concepts and semantic relations between them are – at least to a large extent – language independent. This assumption could however be heavily criticized from a linguistic point of view. It may also pose practical problems during the compilation of new wordnets, given that some of the PWN senses may have no TL counterpart. Consequently, a varying number of TL synsets may be left unfilled (depending on the TL), which limits the usefulness of the newly-built semantic hierarchy in NLP applications.

Other issues posed by this translation-based approach to wordnet development are the manual work needed during transfer and its heavy reliance on external lexico-semantic resources. In EuroWordNet, BalkaNet and MultiWordNet, for instance, PWN literals were mainly translated by human lexicographers using external resources (e.g. dictionaries, thesaurus, taxonomies, etc.).[1] However, the coverage of the external resources used for translating the contents of PWN into new languages may also pose problems during transfer. It may limit the approach to specific language pairs and have a negative impact on the coverage of the newly built resources.

In spite of these theoretical and practical drawbacks, new wordnets still heavily depend on PWN. The methods used for transferring semantic information have however become more or less automatic, limiting the cost of the manual methods employed before. They also exploit lexico-semantic information extracted from parallel corpora, instead of relying on predefined semantic resources. For instance, the French hierarchy WOLF (which will be presented in the next section), was automatically built by exploiting information found in several multilingual resources and parallel corpora (Sagot and Fišer, 2008). Another PWN-based resource for French, the JAWS network, was compiled by combining a bilingual dictionary and syntactic information for disambiguating polysemous nouns (Mouton and de Chalendar, 2010). The multilingual semantic network BabelNet goes a step further by jointly exploiting PWN, Wikipedia and the output of Statistical Machine Translation systems (Navigli and Ponzetto, 2010).

2.2. Data-driven semantic analysis

Another important line of research involves the development of multilingual semantic resources by solely using information coming from corpora, without resorting to PWN. The proposed methods generally exploit translation information found in parallel corpora based on the assumption that the translations of words in real texts offer insights into their semantics (Resnik and Yarowsky, 1997). The Semantic Mirrors method (Dyvik, 1998; Dyvik, 2005), for instance, discovers word senses from parallel corpora, as well as their semantic relations which permit to organize them in a complex lexico-semantic network. Ide et al. (2002) exploit the same assumption for Word Sense Induction (WSI) and use the translations of words in a multilingual parallel corpus as features for building translation vectors. These vectors are clustered and the obtained clusters describe the senses of the source language words.

In the same vein, Van der Plas and Tiedemann (2006) build translation vectors whose similarity reveals the source words' proximity. Apidianaki (2008) combines translation and distributional information for WSI. The translations of the words in a parallel corpus are represented by weighted feature vectors describing the corresponding source language contexts. The vectors and the corresponding translations are clustered according to their similarity, and the acquired sense clusters represent the source word senses. In a semantic annotation setting, Diab and Resnik (2002) combine translation information found in a parallel corpus with semantic information coming from PWN. The possible semantic tags of the English translations of a French word are found in PWN and the one characterizing the whole set of translations is selected and used as the word's sense tag.

In this work, we employ a cross-lingual WSD method for automatically enriching the WOLF with semantic information acquired from a parallel corpus by the WSI method of Apidianaki (2008). Before presenting our method, in Section 3 we provide information on WOLF, the way it was compiled, and its contents and coverage.

3. WOLF

WOLF (Sagot and Fišer, 2008) is a freely available wordnet for French. Its first version (WOLF 0.1.4) was created on the basis of PWN (version 2.0) by following the *expand* model for wordnet development. Monosemous literals in the PWN were translated using a bilingual French-English lexicon built from various multilingual resources.[2] Polysemous PWN literals were handled by an *alignment* approach based on the multilingual parallel corpus SEE-ERA.NET (Steinberger et al., 2006).[3] The corpus was lemmatized, POS-tagged and word aligned, and bilingual lexicons were automatically built including the translations of English words in different languages. These lexicons were then combined into multilingual lexicons and a synset id was assigned to each lexicon entry by gathering all possible ids for this entry in all languages from the corresponding BalkaNet wordnets. The underlying assumption being that it is unlikely that the same polysemy occurs in different languages, the intersection of the possible senses was likely to output only the correct synset. So, the ids shared by all non-French lexicon entries were assigned to their French equivalent. The synsets obtained from both approaches were then merged. The resulting network, WOLF, preserves the hierarchy and structure of PWN 2.0 and contains the definitions and usage examples provided in PWN for each synset.

Compared to the initial version of WOLF (0.1.4), the version used here (0.1.6) has an extended coverage on adverbs, as a result of the work by Sagot et al. (2009).

As information was not found for all PWN synsets by the employed automatic methods, WOLF 0.1.6 is rather sparse. In total, it contains 32,351 non-empty synsets including 37,991 unique literals (vs. 115,424 synsets with 145,627 literals in PWN 2.0). These synsets are filled with 34,827 unique French noun literals, 1,521 adjectives, 979 verbs and 664 adverbs.

The work presented in this paper is aimed at enriching this resource and increasing its coverage. Nevertheless, the proposed methodology can also be applied for developing new wordnet-like resources in other languages on the basis of PWN.

4. Enriching the WOLF

4.1. Cross-lingual WSI and WSD

Filling empty synsets in a wordnet can be achieved by creating clusters of synonyms (synsets) and defining the

[1] In these projects, the *expand* model was also sometimes combined with the *merge* model, which is based on monolingual resources and permits to include language-specific properties in the wordnets of different languages.

[2] Wikipedia, the English and French Wiktionary, Wikispecies and the Eurovoc thesaurus.

[3] The corpus is composed of the English, French, Romanian, Czech and Bulgarian parts of the JRC-Acquis subcorpus.

place where they should be located in the hierarchy. The first task could be carried out by a method capable of identifying the senses of words in texts (i.e., a WSI method), while for the second task a Word Sense Disambiguation (WSD) method would be needed.

For enriching the WOLF, new synsets containing synonymous French words should be acquired from (monolingual French or parallel) text corpora and integrated in the hierarchy. However, the sparsity of the information available in WOLF would hamper the use of a monolingual WSD method for positioning the new synsets in the hierarchy. Given that WOLF has the same structure as PWN 2.0, an alternative would be to exploit PWN information for disambiguating the new synsets. So, the new French synsets could be included in the hierarchy on the basis of information found in the English WordNet, by means of a cross-lingual WSD classifier.

We employ the cross-lingual WSD method proposed by Apidianaki (2009) which is well adapted to the task at hand for several reasons. First, it exploits the results of a WSI method that generates synset-like clusters of the translations of words found in a parallel corpus. The translations are grouped according to their semantic similarity, calculated on the basis of source language distributional information (Apidianaki, 2008). More precisely, the translations are characterized by source language feature vectors whose similarity serves to group the corresponding translations into clusters. When applied to the EN–FR language pair, the method clusters the French translations of English words by comparing the corresponding English feature vectors. The obtained clusters of translations describe the senses of the English words in the corpus and contain semantically close words in French, similar to wordnet synsets. The automatically generated French clusters constitute the synsets to be included in the resource.

The second reason that makes this cross-lingual WSD method well suited for this task is that the proposed WSD classifier selects target language (e.g. French) sense clusters for filling the empty synsets based on source language (e.g. English) information. This is due to the nature of the output of the WSI method (which will be described in more detail in section 4.1). The generated translation clusters are characterized by feature vectors that can be used for assessing the similarity of a cluster and a synset, thanks to information extracted from the PWN (see 4.3).

4.2. Word Sense Induction: creating sense clusters

4.2.1. Training

The WSI method used for acquiring new French synsets is trained on the sentence aligned FR–EN part of the EUROPARL corpus (release v6) (Koehn, 2005).[4] Both sides of the corpus are lemmatized and POS-tagged using TreeTagger (Schmid, 1994), and the corpus is aligned at the level of word types using GIZA++ (Och and Ney, 2003). Two bilingual lexicons are then extracted from the alignment results, one for each translation direction (EN–FR/FR–EN). To eliminate noisy alignments, the translations are filtered on the basis of their alignment score (threshold: 0.01) and according to their POS, keeping for each word (w) only translations pertaining to the same grammatical category. Finally, an intersection filter discards any translation correspondences not found in both lexicons. The translations used for clustering are the ones that translate w more than 10 times in the training corpus. Even if this threshold leaves out some translations of the source words, it has a double merit: it reduces data sparseness issues and eliminates erroneous translations which may be found in the lexicons because of spurious alignments.

4.2.2. Semantic similarity calculation

Each translation of a word w is characterized by a vector built from the lemmas of the content words (nouns, verbs and adjectives) that cooccur with w in the corresponding source language sentences of the parallel corpus. For instance, four vectors are built for the noun *stage* which has four translations in the training corpus (*stade*, *phase*, *étape* and *scène*). Each vector contains the content words that *stage* cooccurs with in the source side of the aligned sentences where it is translated by each French word.

A similarity score is computed for each translation pair by a variation of the Weighted Jaccard measure (Grefenstette, 1994). The input of the similarity calculation for two translations consists of their frequency lists as well as of those built for the other translations of w. The score assigned to a pair of translations indicates their degree of similarity. It is computed as follows.

For each translation T_i of w, each feature F_j ($1 \leq j \leq N$) that occurs in the corresponding source language context receives a *total weight* $tw(F_j, T_i)$. This total weight is defined as the product of the *global weight* of the feature, $gw(F_j)$, and a *local weight* with that translation, noted $lw(F_j, T_i)$. The global weight $gw(F_j)$ is based on the number N_j of translations with which F_j is related and on the dispersion of F_j in the contexts of w. The value of this dispersion relies on the probabilities p_{ij} that F_j cooccurs with instances of w that are translated by each of the T_i's:

$$gw(F_j) = 1 - \frac{\sum_{T_i} p_{ij} \log(p_{T_i F_j})}{N_j} \quad (1)$$

Each of the p_{ij}'s is computed as the ratio between the cooccurrence frequency of F_j with an occurrence of w translated as T_i, noted $\text{cooc\_frequency}(F_j, T_i)$, and the number of features seen with T_i, noted n_i:

$$p_{ij} = \frac{\text{cooc\_frequency}(F_j, T_i)}{n_i} \quad (2)$$

On the other hand, the local weight $lw(F_j, T_i)$ between F_j and T_i directly depends on their frequency of cooccurrence:

$$lw(F_j, T_i) = \log(\text{cooc\_frequency}(F_j, T_i)) \quad (3)$$

Recall now that $tw(F_j, T_i) = gw(F_j) \cdot lw(F_j, T_i)$. The Weighted Jaccard (WJ) coefficient of two translations T_m and T_n is defined as follows:

[4]Sentence pairs with a great difference in length (i.e., where one sentence is more than 3 times longer than the corresponding sentence in the other language) are eliminated.

Language	POS	Source word	Sense clusters
EN–FR	Nouns	omission	{carence}{lacune, oubli ,négligence} {lacune, omission}
		assessment	{analyse, appréciation, bilan, estimation, étude} {évaluation} {jugement, estimation}
	Verbs	accommodate	{adapter, répondre} {satisfaire, répondre} {accueillir}
		combine	{conjuguer, combiner, associer} {réunir, unir, conjuguer, concilier} {fusionner} {conjuguer, concilier, réunir, associer} {ajouter} {regrouper, rassembler, réunir}
	Adjs	dubious	{suspect}{douteux, discutable} {discutable, contestable}
		outstanding	{excellent, suspens, remarquable} {exceptionnel, extraordinaire} {remarquable, exceptionnel, excellent}
FR–EN	Nouns	diffusion	{broadcasting, dissemination, distribution} {circulation} {distribution, diffusion} {broadcasting, distribution, broadcast}
		peine	{sentence, penalty, punishment} {trouble, bother}
	Verbs	menacer	{threaten} {endanger, risk, jeopardise}
		lier	{link, connect, relate} {attach} {combine}
	Adjs	lisible	{comprehensible, legible} {legible, readable}
		malheureux	{sad, unhappy, wretched} {unfortunate}

Table 1: Entries from the sense cluster inventories

$$WJ(T_m,T_n) = \frac{\sum_j \min(tw(T_m,F_j),tw(T_n,F_j))}{\sum_j \max(tw(T_m,F_j),tw(T_n,F_j))} \quad (4)$$

Translation pairs with a score above a threshold defined locally for each *w* are considered as semantically related.[5] The clustering algorithm groups the translations according to their similarity and the obtained sense clusters describe the senses of the corresponding source language words. The clusters generated, for instance, for the English noun *stage* describe its two senses in the training corpus: {*stade, phase, étape*} and {*scène*} (i.e., the "phase" sense and the "platform" sense). The clustering procedure is presented in detail in the next section.

4.2.3. Semantic clustering

The SEMCLU algorithm (Apidianaki, 2008; Apidianaki and He, 2010) groups the translations into clusters by exploiting the similarity calculation results. Its input, for each *w*, consists in: (a) the list of its translations (T_list); (b) their similarity table; (c) the similarity threshold.

The clustering is performed in two steps. First, each translation pair having a similarity score above the threshold is considered to have a pertinent relation and forms an 'initial' cluster (C). These two-element clusters are derived directly from the similarity table. During the second step, they may be enriched by additional translations, by the recursive function 'enrich_cluster' shown in Algorithm 1. The function takes as input the cluster C and the list of translations of *w* (T_list), and outputs C eventually enriched by other translations.

A new translation is included in a cluster if it has pertinent relations with all the elements already in the cluster. The clustering stops when all the translations of *w* are included in some cluster and all their relations have been checked. The final clusters are characterized by *global connectivity*, i.e. all their elements are linked by pertinent relations.

[5]The procedure used for defining the threshold is detailed in Apidianaki and He (2010).

Algorithm 1 *The 'enrich_cluster' function.*

```
enrich_cluster(T_list, C):
if empty T_list then
  return C
else
  if first T in T_list linked to all Ts in C then
    enrich_cluster(rest T_list, C union T)
  else
    enrich_cluster(rest T_list, C)
  end if
end if
```

The translations having no pertinent relation to any other translation of *w* are included in separate one-element clusters.

Two sense cluster inventories are created from the training data. An EN–FR inventory, where the senses of English words are described by clusters of their French translations, and a FR–EN inventory, where the senses of French words are described by clusters of their translations in English. The sense clusters group semantically similar words in the TL and could be compared to wordnet synsets. In Table 1, we present some examples of English and French entries of different POS and degrees of polysemy. The EN verb *accommodate*, for instance, has four translations (*adapter, répondre, satisfaire, accueillir*) which are grouped in three sense clusters: {*adapter, répondre*} ("adapt" sense), {*satisfaire, répondre*} ("satisfy") and {*accueillir*} ("put up"). The two first clusters overlap (they both contain the FR verb *répondre*), which means that the described senses are probably related. The cluster overlaps could actually serve as clues for their merge, if coarser-grained sense descriptions were needed. However, as a translation may be found in the intersection of two clusters because of being ambiguous between the two senses, a merge would be more reliable if the intersection contained more than one element.

In this work, the sense clusters are used for filling FR

POS	EN entry	PWN synset	FR sense cluster
Nouns	**presentation**	ENG20-06725607-n: {presentation#n#4} the act of presenting a proposal	{présentation, exposé}
	scam	ENG20-00709982-n: {scam#n#1, cozenage#n#1} a fraudulent business scheme	{arnaque, escroquerie}
	loyalty	ENG20-04639012-n: {loyalty#n#1} the quality of being loyal	{fidélité, loyauté}
Verbs	**discourage**	ENG20-00841635-v: {warn#v#2, discourage#v#3, admonish#v#1, monish#v#2} admonish or counsel in terms of someone's behavior; ”I warned him not to go too far”; ”I warn you against false assumptions”; ”She warned him to be quiet”	{décourager, dissuader}
	distance	ENG20-02602279-v: {distance#v#1} keep at a distance; ”we have to distance ourselves from these events in order to continue living”	{éloigner, distancier}
	divide	ENG20-02543903-v: {separate#v#1, divide#v#3} act as a barrier between; stand between; ”The mountain range divides the two countries”	{partager, séparer, répartir}
Adjectives	**horrific**	ENG20-01575285-a: {hideous#a#1, horrid#a#2, horrific#a#1, outrageous#a#1} grossly offensive to decency or morality; causing horror; ”subjected to outrageous cruelty”; ”a hideous pattern of injustice”; ”horrific conditions in the mining industry”	{atroce, terrible, épouvantable}
	sure	ENG20-00331475: {indisputable#a#2, sure#a#9} impossible to doubt or dispute; ”indisputable (or sure) proof”	{sûr, certain}
	clever	ENG20-00413048-a: {cagey#a#1, cagy#a#1, canny#a#1, clever#a#2} showing self-interest and shrewdness in dealing with others; ”a cagey lawyer”; ”too clever to be sound”	{habile, astucieux}
Adverbs	**brutally**	ENG20-00204148-b: {viciously#r#1, brutally#r#1, savagely#r#1} in a vicious manner; ”he was viciously attacked”	{brutalement, sauvagement}
	early	ENG20-00101775-b: {early_on#r#1, early#r#1} during an early stage; ”early on in her career”	{rapidement, tôt}
	exactly	ENG20-00372187-b: {precisely#r#2, incisively#r#2, exactly#r#3} in a precise manner; ”she always expressed herself precisely”	{exactement, précisément}

Table 2: Previously empty WOLF synsets filled by our WSD method

synsets that correspond to PWN synsets (i.e., characterized by very fine granularity). Moreover, given that wordnet synsets may in general contain the same literals, the overlaps pose no problem in this context. Consequently, clusters have not been merged but used as proposed by the WSI method.

4.3. Integrating sense clusters into WOLF

The generated EN–FR sense cluster inventory contains entries for English words of different POS. In this paper, as a first experiment, we focus on word meanings corresponding to empty synsets in WOLF. In future work, we intend to enrich non empty synsets as well with additional information found in the clusters.

The unsupervised WSD classifier used (Apidianaki, 2009) exploits the WSI results. In a classic WSD task, the clusters constitute the candidate senses of a word from which the most adequate one has to be selected for each instance of the word in context. This selection is performed by comparing the vectors of the clusters to information in the new context.

In the current setting, where the goal of the WSD method is to assign clusters to empty synsets in WOLF, the information used for WSD consists of the words found in the corresponding PWN synsets and their related synsets, their definitions and usage examples. Given that information in the vectors is lemmatized, the information retained from PWN is lemmatized as well (Schmid, 1994) and gathered in a bag of words. The adequacy of a cluster for filling a given synset is estimated by comparing the cluster's vector with the PWN information retained for the synset. If common features (CFs) are found with just one cluster, this cluster is selected. Otherwise, each 'cluster-synset' association is assigned a score corresponding to the mean of the weights of the CFs relatively to the clustered translations (weights assigned to each feature during WSI (cf. 4.2.)). In formula 5, the CF_j is the set of CFs found between the cluster and the synset and N_{CF} is the number of translations T_i in the cluster characterized by a CF. The highest scored cluster is selected and assigned to the empty synset.

$$assoc\_score = \frac{\sum_{i=1}^{N_{CF}} \sum_j w(T_i, CF_j)}{N_{CF} \cdot |CF_j|} \quad (5)$$

PWN entry *peaceful* (adj)

French cluster	The English vector of the cluster represented as a bag of words
{*paisible*, *pacifique*}	***absence acceptance*** *achieve action* ***activity aggressive*** *agreement* ***atmosphere attitude*** *authority* ***be become*** *believe bring* ***call calm*** *can* ***citizen*** *clear coexistence commission community conflict continue cooperation council* ***country*** *crisis democracy democratic demonstration demonstrator development dialogue* ***dispute*** *do east economic effort election* ***emotional*** *energy ensure* …

Corresponding PWN synsets	Synset-related information, represented as a bag of words
ENG20-01615936-a {*law-abiding*, *peaceful*} Def.: *(of groups) not violent or disorderly* Usage: *the right of peaceful assembly* Neighboring synsets: ENG20-01615787-a {*orderly*}	*an assembly confront crowd devoid disorderly disruption group law-abiding not of or* ***orderly*** *peaceful* ***president right*** *the* ***violence*** *violent*
ENG20-01686906-a {*peaceful*} Def.: *not disturbed by strife or turmoil or war* Usage: *a peaceful nation; peaceful times; a far from peaceful Christmas; peaceful sleep* Neighboring synsets: ENG20-01687344-a {*calm, serene, tranquil*} ENG20-00302191-a {*calm*} ENG20-01202829-a {*amicable*} …	*a* ***absence*** *abstain* ***acceptance activity aggressive*** *agitation almost amicable an and antagonist assertiveness at* ***atmosphere attitude be become*** *by* ***call calm*** *characterize* ***citizen*** *conducive* ***country*** *directly dispose* ***dispute*** *disturb disturbance dovish* ***emotional*** …
ENG20-02425529-a {*passive*, *peaceful*} Def.: *peacefully resistant in response to injustice* Usage: *passive resistance* Neighboring synsets: ENG20-02425368-a {*nonviolent*}	*abstain from in injustice nonviolent of on* ***passive*** *peaceful peacefully principle resistance resistant response the to use violence*

Cluster-to-synset mapping: the bag of words representing the synset ENG20-01686906-a is the closest to that of the vector of the French cluster {*paisible*, *pacifique*}, and it also gets the highest score during WSD.[6]

Outcome: *paisible* and *pacifique* are added to synset ENG20-01686906-a in the WOLF

Figure 1: Comparison of vector and PWN information during WSD

For instance, the empty synset 'odd#a#2' (definition: "not easily explained"; usage: "it is odd that his name is never mentioned"), is correctly filled by the FR cluster {*curieux*, *bizarre*}. The other clusters of *odd,* which were scored less, are: {*contradictoire*, *singulier*, *bizarre*} and {*curieux*, *étrange*}. More examples of synsets filled by the WSD method are shown in Table 2. We provide the PWN id of the empty synsets; the EN headword; the literals that the corresponding synsets contain in PWN, as well as their definition and usage examples.[7] The French literals in the sense cluster most strongly associated with a PWN synset, which are used to fill the corresponding synset in WOLF, are given in the last column of Table 2.

The process of selecting the French synset that best suits a cluster on the basis of English contextual information is illustrated with the example given in Figure 1. It details the case for the English adjective *peaceful*, which belongs to three synsets in the PWN that are all empty in WOLF. The WSD method has to fill one of these synsets with the cluster {*paisible*, *pacifique*} associated with *peaceful*. Each of the PWN synsets for *peaceful* is shown in Figure 1 (literals, definition, usage examples), together with (some of) the related synsets that are used to build their corresponding bags of words. The words that belong at the same time to the bag of words created from the cluster vector and to the bag of words of one of the synsets are shown in boldface. The bag of words representing the synset ENG20-01686906-a is the closest to that of the vector of the French cluster, and it also gets the highest score during WSD. Therefore, *paisible* and *pacifique* are added to synset ENG20-01686906-a in the WOLF.

[6]The weights of the features are not given here, for the sake of readability.

[7]This table does not include information on all the neighboring PWN synsets (which was also used during WSD). This information can however be easily recovered from PWN.

5. Evaluation

Overall, 3,904 previously empty synsets have been filled by our approach (2,333 nominal, 576 verbal, 709 adjectival and 286 adverbial synsets). We have manually examined 10% of the synsets filled for each POS, for evaluating the quality of the proposed clusters and the correctness of their assignment to some synset in WOLF. A cluster is considered as a good quality one if it groups words that share the same meaning. The assignment of a cluster to a synset is considered as correct if its contents correctly describe the sense in the corresponding PWN synset. Of course, a cluster may be correctly assigned to a synset only if it is of good quality according to the first evaluation criterion.

Both aspects have been evaluated by two annotators. The inter-annotator agreement was measured at $\kappa = 0.67$ for cluster quality, and 0.59 for the WSD results, which is conventionally interpreted as "good" agreement (Cohen, 1960).

According to the evaluation results obtained for all POS, the clusters group semantically similar words in 75.5% of the cases. Significant variations are however observed for different POS, as shown in Table 3. The first row of the table contains the percentage of good quality clusters in the test set. The second row shows the percentage of the clusters that were correctly assigned to WOLF synsets.

The observed divergences are due to the restrictive cluster quality criterion used, according to which one incorrect word in an otherwise correct cluster turns the whole cluster into an incorrect one. This strict criterion unfairly penalizes and rejects interesting although noisy clusters. We notice that this criterion has a strong effect especially on clusters containing many translations, as is often the case for verb clusters. We plan to proceed to a more detailed and flexible evaluation to more accurately estimate the actual merit of the clustering method. This will also imply devising methods for cleaning noisy clusters .

	Nouns	**Verbs**	**Adjs**	**Advs**
Clusters	72.1	62.9	81.0	86.2
WSD	64.6	53.0	75.1	73.7

Table 3: Evaluation Results (%)

The noise found in the clusters may be due to alignment errors that were not eliminated by the filters used to clean the lexicons (cf. 4.2.), or it may be introduced during the clustering procedure. The error analysis indicates some cases of problematic clustering that fall into the second category.

(a) cases where multiword units were not considered during word alignment. This is observed in the cluster {*considération*, *compte*} corresponding to *consideration*, which should ideally be {*prise en compte*, *considération*}

(b) clustering of topically related but not synonymous words, as in the cluster {*raisin*, *moût*} corresponding to *grape*

(c) clustering of antonymous but distributionally similar words, as in the case of {*sain*, *malsain*} (cluster of *unhealthy*). Antonymous words may be found in the alignment results when the negation is expressed paraphrastically in one of the languages (e.g., French) and is not captured by the alignment, as is here the case with the translation *sain* retained for *unhealthy*. Then, as antonymous words often appear in similar contexts, it happens that they end up in the same cluster.

Given that only good clusters can be correctly integrated into WOLF, we calculate the performance of the WSD method by reference to the number of good clusters. The score obtained for the WSD insertions by averaging the scores provided by the two annotators is 67%, which is very encouraging. We should highlight the difficulty of this task as the WSD method is asked to fill synsets that were left empty by the methods initially employed for creating WOLF. These empty synsets often correspond to rare senses in PWN, that may not exist in the training corpus, or to senses for which little information is available.

In order to more fairly estimate the performance of the WSD method in this setting, we also tested it on the whole resource. In this case, the method was asked to select the most appropriate synset for each cluster from *all* the synsets in WOLF (not only the empty ones). In this setting the method reaches a performance of 80.13%, which shows that it is particularly well adapted to the wordnet development task.

6. Conclusion

We have shown that a cross-lingual WSD method based on unsupervised Word Sense Induction can be efficiently used for wordnet development. We integrated sense clusters of translations into a French wordnet resource, the WOLF, by exploiting information found in PWN. The results indicate that the proposed unsupervised methods are particularly useful for the construction and enrichment of wordnets in languages other than English.

We conclude with some issues for future research. Based on these encouraging results, we intend to use the proposed methods in order to enrich other, non-empty, synsets in WOLF. Moreover, we will apply measures of semantic similarity on PWN in order to merge closely related synsets and, consequently, reduce the number of empty ones. For this we will also exploit the cross-lingual WSD results, in particular in cases where a cluster is selected as adequate for filling different synsets. This will serve as a clue for automatically estimating the similarity of synsets, merging them into coarser-grained ones and further reducing the sparseness of the resource.

Acknowledgements

This work was funded by the French ANR national grant EDyLex (ANR-09-CORD-008).

7. References

Marianna Apidianaki and Yifan He. 2010. An algorithm for cross-lingual sense clustering tested in a MT evaluation setting. In *Proceedings of the 7th International*

Workshop on Spoken Language Translation (IWSLT), pages 219–226, Paris, France.

Marianna Apidianaki. 2008. Translation-oriented sense induction based on parallel corpora. In *Language Resources and Evaluation Conference (LREC)*, pages 3269–3275, Marrakech, Morocco.

Marianna Apidianaki. 2009. Data-driven semantic analysis for multilingual WSD and lexical selection in translation. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, pages 77–85, Athens, Greece.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46.

Mona Diab and Philip Resnik. 2002. An Unsupervised Method for Word Sense Tagging using Parallel Corpora. In *Proceedings of the the 40th Annual Meeting of the Association for Computational Linguistics (ACL'02)*, pages 255–262, Philadelphia.

Helge Dyvik. 1998. Translations as semantic mirrors: from parallel corpus to wordnet. In *Proceedings of the Workshop Multilinguality in the lexicon II at the 13th biennial European Conference on Artificial Intelligence (ECAI'98)*, pages 24–44, Brighton, UK.

Helge Dyvik. 2005. Translations as a semantic knowledge source. In *Proceedings of the Second Baltic Conference on Human Language Technologies*, Tallinn, Estonia.

Christiane Fellbaum, editor. 1998. *WordNet: An Electronic Lexical Database*. MIT Press, Cambridge, Massachusetts.

Gregory Grefenstette. 1994. *Explorations in Automatic Thesaurus Discovery*. Dordrecht : Kluwer Academic Publisher.

Nancy Ide, Tomaž Erjavec, and Dan Tufiş. 2002. Sense discrimination with parallel corpora. In *Proceedings of ACL'02 Workshop on Word Sense Disambiguation: Recent Successes and Future Directions*, pages 54–60, Philadelphia.

Philipp Koehn. 2005. Europarl: A Parallel Corpus for Statistical Machine Translation. In *Proceedings of MT Summit X*, pages 79–86, Phuket, Thailand.

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

Roberto Navigli and Simone Paolo Ponzetto. 2010. BabelNet: Building a Very Large Multilingual Semantic Network. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 216–225, Uppsala, Sweden.

Franz Josef Och and Hermann Ney. 2003. A Systematic Comparison of Various Statistical Alignment Models. *Computational Linguistics*, 29(1):19–51.

Emanuele Pianta, Luisa Bentivogli, and Christian Girardi. 2002. Multiwordnet: developing an aligned multilingual database. In *First International Conference on Global WordNet*, pages 293–302, Mysore, India.

Philip Resnik and David Yarowsky. 1997. A perspective on Word Sense Disambiguation Methods and their Evaluation. In *Proceedings of the ACL SIGLEX Workshop on Tagging Text with Lexical Semantics : Why, What, and How?*, Washington, D.C.

Benoît Sagot and Darja Fišer. 2008. Building a free french wordnet from multilingual resources. In *Ontolex 2008*, Marrakech, Morocco.

Benoît Sagot, Karën Fort, and Fabienne Venant. 2009. Extending the adverbial coverage of a french wordnet. In *Proceedings of the NODALIDA 2009 workshop on WordNets and other Lexical Semantic Resources*, Odense, Danemark.

Helmut Schmid. 1994. Probabilistic Part-of-Speech Tagging Using Decision Trees. In *Proceedings of the International Conference on New Methods in Language Processing*, pages 44–49, Manchester, UK.

Ralf Steinberger, Bruno Pouliquen, Anna Widiger, Camelia Ignat, Tomaž Erjavec, Dan Tufiş, and Daniel Varga. 2006. The JRC Acquis : A multilingual aligned parallel corpus with 20+ languages. In *Proceedings of LREC*, Genoa, Italy.

Dan Tufiş, Dan Cristea, and Sofia Stamou. 2004. BalkaNet: Aims, Methods, Results and Perspectives. A General Overview. In *Romanian Journal on Information Science and Technology. Special Issue on BalkaNet*, volume 7, pages 9–34.

Lonneke Van der Plas and Jörg Tiedemann. 2006. Finding synonyms using automatic word alignment and measures of distributional similarity. In *Proceedings of ACL/COLING 2006*, pages 866–873, Sydney, Australia.

Piek Vossen, editor. 1998. *EuroWordNet: a multilingual database with lexical semantic networks for European Languages*. Kluwer, Dordrecht.