FrameNet translation using bilingual dictionaries with evaluation on the English-French pair

Claire Mouton[12], **Gaël de Chalendar**[1], **Benoit Richert**[1]

[1] CEA, LIST, Laboratoire Vision et Ingénierie des Contenus, Fontenay aux Roses, F-92265, France;
[2] Exalead S.A. Paris, France
Claire.Mouton@cea.fr,Gael.de-Chalendar@cea.fr

Abstract

Semantic Role Labeling cannot be performed without an associated linguistic resource. A key resource for such a task is the FrameNet resource based on Fillmore's theory of frame semantics. Like many linguistic resources, FrameNet has been built by English native speakers for the English language. To overcome the lack of such resources in other languages, we propose a new approach to FrameNet translation by using bilingual dictionaries and filtering the wrong translations. We define six scores to filter, based on translation redundancy and FrameNet structure. We also present our work on the enrichment of the obtained resource with nouns. This enrichment uses semantic spaces built on syntactical dependencies and a multi-represented k-NN classifier. We evaluate both the tasks on the French language over a subset of ten frames and show improved results compared to the existing French FrameNet. Our final resource contains 15,132 associations *lexical units-frames* for an estimated precision of 86%.

1. Introduction

Text semantic analysis aims at providing a machine the information it requires to perform complex tasks such as natural language question analysis or reasoning. One approach to semantic analysis is the *Semantic Role Labeling* (*SRL*). Several semantic resources have been built describing sets of standard situations and associated sets of roles that text phrases can fill.
VerbNet (Schuler, 2005) lists a global set of verbal roles and attributes to each verb a subset of them. VerbNet contains currently more than 3,700 verbs. Propbank (Kingsbury and Palmer, 2002) is similar to VerbNet but roles are less semantically typed. Berkeley project FrameNet (Baker et al., 1998) contains a set of frames, each describing a precise situation that can appear in our perception of the world. Each frame is described with a set of roles specific to the frame. FrameNet also lists predicates (verbs, nouns, adjectives or even adverbs) called Lexical Units (*LU*) triggering these situations. The database contains more that 10 000 *LUs*. *Semantic Role Labeling* consists in attributing semantic roles described in these resources to text phrases. Table 1 quickly compares these resources. French resources for *SRL* are rare. Volem (Fernandez et al., 2002) is a manually built 1,500 verbs database while (Padó and Pitel, 2007) produced a FrameNet automatic translation based on the method firstly proposed in (Padó and Lapata, 2005). They use the alignment of annotated parallel corpora to generate French *LUs* from their alignment with English *LUs*. As a reference, we will call this resource FrameNet.Fr in the remaining of this paper. (Padó and Pitel, 2007) compared their approach with one using a bilingual dictionary, the quality of the latter being worse. In this work, we propose a new bilingual dictionaries approach. We filter the translation pairs obtained by making use of the polysemy made explicit in dictionaries. Our approach can use any bilingual dictionary making distinctions between senses or not. It is theoretically transposable to any language but we must admit that half our study is based on the French part of the Wiktionary[1], which is the most developed one[2]. In 2005 it was filled automatically on the base of free dictionaries and the community is very active to manually enrich the resource ever since. In this work, we evaluate only the French translation.

Resource	Main roles	Sentence Example
VerbNet	21	He *[agent]* left *[keep-15.2]* the car *[theme]* in the park *[location]*
PropBank	5	He *[Arg0]* left *[leave.02]* the car *[Arg1]* in the park *[Arg2]*
FrameNet	250	He *[theme]* left *[Departing]* the car *[source]* in the park *[place]*

Table 1: Semantic roles in different resources

[1] http://wiktionary.org

[2] On March 15th, 2010: French part: + 1,659,000 entries, English part: + 1,609,000.
Only 18 languages contained more than 100,000 entries (in decreasing order): French, English, Lithuanian, Turkish, Chinese, Russian, Vietnamese, Ido (coming from Esperanto), Polish, Portuguese, Finnish, Hungarian, Norwegian, Tamil, German, Italian, Swedish, Greek.
We can also notice that the German part contains only 104,294 entries but represents 12.8% of the traffic versus 9.7% for the French part !

Some work related to the present one also made use of Wiktionary as a bilingual resource with the purpose of translating English resources (WordNet) into French ((Sagot and Fišer, 2008)).

As the original Berkeley resource is manually built and not yet exhaustive, we also study the possible enrichment of our produced resource. The translation method can deal with *LUs* of the target language for which English translations are already listed as *LU* in the original FrameNet. However, it does not enable to categorize target *LUs* to frames when English equivalents are not in FrameNet yet.

Using a classifier would allow to attribute new *LUs* to existing frames and to reinforce the attributions already made by the translation method. (Pennacchiotti et al., 2008) proposed two different measures of similarity to perform such a task: first, using similarities obtained from three vector spaces types (i.e.window co-occurrences, syntactic co-occurrences and untyped syntactic co-occurrences), and secondly using similarities obtained from WordNet.

We reuse this distributional approach with the multi-represented spaces described in (Grefenstette, 2007) and (Mouton et al., 2009) by using a multi-represented KNN classifier proposed by (Kriegel et al., 2005).

Section 2. describes our translation pairs extraction method. Section 3. presents the filtering heuristics. The extraction and the filtering are evaluated in section 4. The resource enrichment is depicted in section 5. We conclude and propose future works in section 6.

2. Extraction of translation pairs from bilingual lexicographic resources

The FrameNet paradigm is based on Fillmore's semantic frames theory (Fillmore, 2006). Frames contain triggering words called Lexical Units and a set of semantic roles. For example, the frame `Ingestion` includes roles such as `Ingestor`, `Ingestibles`, `Degree`, `Duration`, `Instrument`, and *LUs* `breakfast.v`, `consume.v`, `dine.v`, `drink.v`, `sip.v`, `sip.n`, etc. In a first assumption, we consider frames and roles as purely semantic and independent of the language. We keep the original structure of FrameNet to transpose it to French.

We use two different bilingual dictionaries to translate FrameNet lexical units, namely the community-based Wiktionary and *SCI-FRAN-EuRADic*, a dictionary realized and evaluated by linguists in the framework of the EuRADic project and distributed by the ELDA[3] organization.

The Wiktionary is a set (one by edition language) of multilingual collaborative dictionaries, hosted by the Wikimedia foundation. Their main advantage over traditional dictionaries is inherent to their nature. They are filled every day by their users so that neologisms will quickly be reported. We work with a version dated January 20th, 2009 for the French Wiktionary and February 3rd, 2009 for the English one. At that time, the French resource contained 1,194,408 pages and the English one 1,209,371.

Both EuRADic and Wiktionary handle multiword expressions which was also the case in the approach of (Padó and Lapata, 2005). Moreover, Wiktionaries share one characteristic we want to emphasize: dictionary entries often categorize translations depending on distinct meanings of the word. We will see in the next section how this feature can help filtering the data.

Concerning the English to French translation, and consequently using only English *LUs* present in FrameNet, we obtain 19,912 pairs *French LUs-frame* issued from 27,109 translation pairs with the Wiktionary and 57,787 pairs *Franch LUs-frame* with the help of EuRADic.

3. How to filter: definition of scores

LUs obtained in the target language are not perfect due mainly to polysemy of the English words. For instance, the English *LU depression* present in the frames *Medical_conditions* and *Natural_features* is translated to *dépression*, *abattement* and *mélancolie*. *dépression* is affected correctly to the *Natural_features* and *Medical_conditions* frames, but *abattement* and *mélancolie* are also affected to the same frames, which is right for *Medical_conditions* but does not hold for the frame *Natural_features*.

Aiming to keep only the correct translations, we associate scores to each *LU-Frame* pair. *LUs* whose scores are higher than a parameter threshold will be kept. Firstly, we define an initial score S1 from which we compute five other scores Si based on different heuristics. The two first of them make use of the structure of the source Berkeley FrameNet while the last ones make use of the target data produced by translation.

[3]ELDA - Evaluations and Language resources Distribution Agency: http://www.elda.org/

3.1. S1 score

For a given target *LU*, S1 is the sum of the number of dictionary senses of all the English *LUs* of the current frame that have this target *LU* as a translation. In our case, the number of dictionary senses is always 1 with EuRADIC while it can vary when using the Wiktionary (since translations are classified by sense).

remettre.v {put back.v: 1} 1
boire.v {quaff.v: 1, drink.v: 2} 3
alimenter.v {feed.v: 1} 1
déjeuner.v {lunch.v: 1, dine.v: 1,
feed.v: 1, eat.v: 1} 4
(...)

Table 2: Sample of scores for the frame *Ingestion*

E.g. from the frame *Ingestion* we produce the pairs *drink.v - boire.v* and *quaff.v-boire.v*. *drink.v - boire.v* is scored 2, because *boire* is a translation of *drink* for the senses *consume liquid through the mouth* and *consume alcoholic beverages*. *quaff.v-boire.v* is scored 1 as it has only one sense in the Wiktionary. Thus, S1 score of *boire.v* is 3.

Wiktionary language	Wiktionary entry	Source language	Target language
French	boire	French	English
French	drink	English	French
English	boire	French	English
English	drink	English	French

Table 3: Four locations for translations in Wiktionary

Due to the structure of Wiktionaries, there is four distinct possibilities to extract translation pairs as shown in table 3. When a translation pair appears in several Wiktionary locations, we use the maximal score produced by this pair in one of the locations, considering it corresponds more to the possible number of senses than the sum of scores would. For example *boire.v* is a translation of *drink.v* with a score of 2 in the English Wiktionary (*drink* page), because this translation is present for the two distinct senses *consume liquid through the mouth* and *consume alcoholic beverages*. Its global score is thus 2 even if it has a score of 1 in the three other locations, because the two senses have not been split there.

3.2. Structural scores

If a source word is polysemic (i.e. present in more than one frame), there is more risk that the translation fails. Therefore we introduce the S2 score requiring that the initial score for these specific words has to be higher in order not to be filtered.

$$S2 = \frac{S1}{Number\_of\_frames\_containing\_the\_source\_LUs^{\alpha}}$$

The score of each translated *LU* is divided by the number of frames containing the source *LUs*. This number is elevated to power α, which allows to modulate the impact of the filter on the score.

Let us take the English *LU rise* as an example. It is present in various frames, including the frame *Getting_up*. Two of its translations are *augmenter* and *se lever*. As we increase the requirement because *rise* appears in many frames, its translations need to have a higher S1 score to be kept after filtering. This is the case for the *LU se lever* which is also a translation of the English *get up* appearing in the eponym frame. Having no other English translation in this frame, *augmenter* will be filtered.

The score S1 tells us how much a translation is reliable in a given frame. The more source *LUs* the given frame has, the more probably the S1 scores of the translations will be incremented. We can then become less tolerant to low scored target *LUs*. This is realized by score S3:

$$S3 = \frac{S1}{Number\_of\_source\_LUs\_in\_the\_frame^{\alpha}}$$

As in S2, an α coefficient allows to modulate the filter power.

For instance, the frame *Container* has 119 English *LU*. It is therefore much easier to give a high S1 score to a French *LUs* of this frame, like *bac* which appears to be the translation of 15 of the 119 English *LUs*[4]. With the score S3, we become more demanding with *LUs* in frames containing a high number of source *LUs*.

3.3. Target scores

By going further with the last idea developed in 3.2., we say that each translation favors the incrementation of the S1 score, especially when one source *LU* produces many translations. We can then consider not the number of source *LUs* anymore, but the number of translations produced in the given frame. That is how we define score S4:

$$S4 = \frac{S1}{Number\_of\_translation\_pairs\_in\_the\_frame^{\alpha}}$$

We can also consider that the less *LUs* a frame has, the more an individual *LU* is important for this frame.

[4]jar, bucket, chest, bottle, case, can, pail, urn, container, crate, sack, pot, tin, box, jug

The S5 score goal is to reduce the importance of *LUs* present in large frames. As a consequence, in large frames we only keep the best scored *LUs* and get rid of erroneous ones without worrying too much about false negative. Whereas in small frames, we tend to be more indulgent at the expense of precision. S5 is defined as:

$$S5 = \frac{S1}{Number\_of\_target\_LUs\_in\_the\_frame^{\alpha}}$$

The score of each translated *LU* is divided by the number of *LUs* translated in the given frame. If we apply this filter with $\alpha = 1$ on our example. The *Ingestion* frame has 47 translated *LUs*. The initial score of *boire.v* is $S1 = 3$. Thus, its S5 score is $S5 = \frac{3}{47} = 0,064$.

The more a *LU* is present in a lot of frames, the less it conveys meaning in a given frame. The S6 score tries to take this idea into account by decreasing scores for *LUs* appearing in several frames. S6 is computed as:

$$S6 = \frac{S1}{Number\_of\_frames\_containing\_the\_target\_LUs^{\alpha}}$$

The score of each target *LU* is divided by its number of occurrences in all the frames. We take here as an example the *LU rue.n*, whose S1 score is 1 in the *Roadways* frame (translated from *street*) and 1 also in the *Measure_linear_extent* frame (translated from *block*). Thus in the *Roadways* frame, $S6 = \frac{1}{2} = 0.5$.

4. Evaluation

4.1. Evaluation criteria

The indicators we are interested in are very common: $Precision$, $Recall$, $F - measure$ and $F_{0.5} - measure$. The weight of P and R are equal for F, while precision is favored for $F_{0.5}$. Here follow the respective formulas:

$$P = \frac{Number\_of\_correct\_LUs}{Number\_of\_present\_LUs}$$

$$R = \frac{Number\_of\_correct\_LUs}{Number\_of\_correct\_LUs\_in\_Gold\_Standard}$$

$$F_{\beta} = (1 + \beta^2)\frac{PR}{\beta^2 P + R} \qquad (F = F_1)$$

Measures are computed for each evaluated frame and averaged. This allows to give an equal importance to all frames independently of their size.

4.2. Development and test set

As these scores contained parameters to be set, we produce both a development set to optimize the parameters and a test set to evaluate the final productions. Both sets are built by making a global resource resulting from the union of the three following resources: FrameNet.fr, the unfiltered resource based on direct translation using the Wiktionary and the unfiltered resource based on direct translation using EuRADic.

For the development set, we choose a sample of 10 frames such that their number of *LUs* is representative of the global distribution (quantiles). The corresponding frames of the global resource are corrected by manually removing *LUs* judged incorrect. The resulting set becomes a gold standard with which the given frames of every produced resource will be compared.

As regards the test set, we select the 10 frames used by (Padó and Pitel, 2007). The same process as for the development set is applied to the global resource in order to build the test set.

4.3. Filters parameter setting

By parameters setting, we can produce resources with different properties. We try to build a resource of reasonable size while keeping a rather good precision. In order to obtain such a result, we maximized the $F_{0.5} - measure$, which favors precision at the expense of recall. We prefer to emphasize precision more than recall, since the number of *LUs* in the raw translated resource is very high compared to the original FrameNet. Parameters we make vary are: α parameters and a threshold below which *LUs* are eliminated since considered not reliable.

Resource	α	P	R	$F_{0.5}$
Wiktio		63%	40%	53%
Wiktio+S1 $F_{0.5}max$		63%	40%	53%
Wiktio+S2 $F_{0.5}max$	1	65%	**40%**	54%
Wiktio+S3 $F_{0.5}max$	1	63%	40%	53%
Wiktio+S4 $F_{0.5}max$	0.5	66%	38%	53%
Wiktio+S5 $F_{0.5}max$	0.75	66%	38%	53%
Wiktio+S6 $F_{0.5}max$	1	70%	36%	**55%**
EuRADic		51%	93%	56%
EuRA+S1 $F_{0.5}max$		74%	34%	58%
EuRA+S2 $F_{0.5}max$	0.75	59%	**75%**	60%
EuRA+S3 $F_{0.5}max$	0.25	69%	51%	59%
EuRA+S4 $F_{0.5}max$	0.1	71%	46%	60%
EuRA+S5 $F_{0.5}max$	0.25	71%	46%	60%
EuRA+S6 $F_{0.5}max$	0.25	68%	55%	**64%**

Table 4: Filters parameter settings on the dev. set

Scores, reported in Table 4, behave very differently depending on the dictionary used. This can be explained by the fact that EuRADic does not contain any sense distinction. All the translation pairs are con-

sequently scored to 1, which is not the case with the Wiktionary. We can however notice a few interesting points: *S2* provides the best recall and *S6* provides the best $F_{0.5}$ in both cases.

4.4. Filters combination

Now that the α parameters have been optimized, we can fix them to combine the filters together. To maximize the filters effect we chose to linearly combine them. For this purpose, scores are normalized (deviation 1, average 0).

$$Score = \sum_{i=\in[|1..6|]} w_i.Si$$
$$with\ w_i \in \{0, \frac{1}{4}, \frac{1}{2}, \frac{3}{4}, 1\}\ and \sum_i w_i = 1$$

We proceed to a systematic variation of the weights of the linear combination and of the filtering threshold to produce the resource with the best $F_{0.5}$ on the development set and fix the weights w_i. We also construct a robust resource with a precision around 95% ($P_{0.95}$). We keep the parameters resulting in the best recall for a precision arbitrarily fixed in the range $[0.945, 0.955]$.

4.5. Results Analysis

The results reported in Table 5 are calculated on the 10 frames test set used by (Padó and Pitel, 2007) in their own study. We found many interesting results: resources obtained by maximizing $F_{0.5}$ have a reasonable precision (74% for both the Wiktionary-based resource and the EuRADic-based one, against 77% for FrameNet.fr) while offering already more *LU-frame* pairs than the Berkeley FrameNet.

For both types of resources, each defined score played a significant role in the filtering, except the score S3. We also made the run for $F_{0.25}max$: it appeared that the best α parameters and weights of combination were slightly different. We noticed also that this time S3 was used in some of the best combinations. These results show that we have to train the score parameters and the linear combination weights each time we want to produce a resource with different characteristics, depending on the task it will be used for.

We also show that we can obtain a 95% estimated precision result for a FrameNet covering 724 of the 796 frames of Berkeley FrameNet while combining the Wiktionary- and EuRADic-based resources (the former attempt FrameNet.fr only reached 480 of them for a lower estimated precision). The last line of the table is added for information and presents a result produced by joining all the intersections of two resources ($Wi\_F_{0.5}max \cap EuRADic\_F_{0.5}max$, $Wi\_F_{0.5}max \cap FrameNet.fr\_nofilter$ and $FrameNet.fr\_nofilter \cap EuRADic\_F_{0.5}max$).

Compared to independent scores, filtering proved to be more efficient when scores are combined and results show that the six of them are useful to increase the precision of the translated resources.

Finally, we notice that the gain in precision is higher for the EuRADic-based resource (58% to 74%), but it is made at the expense of the size of the resource: $Eu\_F_{0.5}$ is reduced to less than the half of its original size while $Wi\_F_{0.5}$ keeps around 80% of its own original size. As EuRADic is much bigger than the Wiktionary concerning the translation part, EuRADic still enables to produce a bigger resource. However, the difference in the reduction size ratios lets us believe that filters are more adapted to the structure of the Wiktionary and that a Wiktionary containing as many entries as EuRADic would help produce a better resource than the one based on EuRADic.

5. Enrichment of the French lexical resource

The fact that our resources can be three times bigger than the original Berkeley FrameNet ($W\_F_{0.5max} \cup E\_F_{0.5max}$) shows that the original English resource is not exhaustive. For example, some *LUs* should be included but appear in no frame (e.g. *taxonomy.n* should appear in the *LUs* of *Categorization*) while others are present but do not belong to all the frames they should trigger (e.g. *boom.n* appears only in *Sounds* while it should be present also in *Progress*). Their translations are therefore generally not present in the relevant frames either. We address this issue by enriching our French resource.

5.1. Enrichment using multi-represented k-NN

To enrich our French resource, we apply a classifier on all the nouns of the vocabulary. The classes to which it affects these words are the different frames consisting of *LUs* obtained by translation.

If the word is already in the resource it is excluded from the frames to which it belongs and reassigned to either confirm a first attribution obtained by translation or produce a new association frame/LU not present originally in FrameNet.

We have at our disposal several semantic spaces computed on syntactical cooccurrences found in a French corpus of over 2 millions web pages. To each syntactical relation (*subject, direct object, noun complement, subject attribute* among others...) detected by

	Linear combination	All frames		Test set			
Resource		#LU-frames	#Frames	P	R	F_{05}	F
Berkeley FrameNet		**11,171**	**796**				
Wi_nofilter		19,912	781	70%	33%	57%	44%
Wi_P095	$\frac{1}{4}.S2 + \frac{1}{4}.S5 + \frac{1}{2}.S6$	2,889	686	94%	11%	33%	18%
Wi_F05max	$\frac{1}{4}.S1 + \frac{1}{2}.S4 + \frac{1}{4}.S6$	**15,720**	**781**	**74%**	**30%**	**56%**	**42%**
EuRADic_nofilter		57,787	795	58%	84%	61%	67%
EuRADic_P095	$\frac{3}{4}.S2 + \frac{1}{4}.S6$	616	210	100%	2%	10%	4%
EuRADic_F05max	$\frac{1}{4}.S2 + \frac{3}{4}.S6$	**24,885**	**767**	**74%**	**44%**	**63%**	**53%**
FrameNet.fr_nofilter		**6,659**	**480**	**77%**	**23%**	**43%**	**31%**
Union							
$Wi \cup EuRADic$		65,488	796	57%	92%	61%	69%
$W\_P_{0.95} \cup E\_P_{0.95}$		3,256	695	94%	12%	35%	20%
$W\_F_{0.5max} \cup E\_F_{0.5max}$		**34,121**	**793**	**70%**	**59%**	**67%**	**63%**
Intersection							
$Wi \cap EuRADic$		**12,211**	**773**	**82%**	**25%**	**56%**	**38%**
$W\_F_{0.5max} \cap E\_F_{0.5max}$		**6,484**	**724**	**95%**	**15%**	**43%**	**25%**
$Wi\_F_{0.5}max \cap Eu\_F_{0.5}max \cap FN.fr$		7,814	742	95%	18%	49%	29 %

Table 5: Evaluation of different sources on the test set

the syntactic analyzer of (Besançon and de Chalendar, 2005), corresponds one semantic space. More details are presented in (Grefenstette, 2007). For each word, we have as many representations as semantic spaces, and we showed in (Mouton et al., 2009) that each of them carry different information. Therefore we apply a classification algorithm proposed by (Kriegel et al., 2005), which is specific to multi-represented data.

This algorithm is a variant of the traditional k-NN classifier. It combines the k-NN spheres of all the representations while taking into account their quality. The idea is that a k-NN sphere with few classes and a lot of elements in each of them is of better confidence than a k-NN sphere with a lot of classes containing few elements. Let R_i be a set of representations and C_j a set of classes, the multi-represented classification rule defined is the following:

$$Cl_{mr}(o) = max_{j=1,...,|C|}(\sum_{i=1}^{m} w_i.cv_{i,j}(o))$$

$cv_{i,j}$ represents the confidence we can have for a class j in a representation i. We will see more details about it in the next subsection of the paper.

w_i is a term corresponding to the entropy with respect to all possible classes. It depends on $cv(o)$. We encourage the reader to refer to (Kriegel et al., 2005) for more theoretical details.

5.2. Tuning parameters

Many parameters have to be set in order to produce the enriched resource.

5.2.1. Learning data

We train the classifier with three different set of data: the best $F_{0,5}$ scored resource (*F*), the biggest resource with precision at $P_{0.95}$ (*P*), and the unfiltered union of the three resources (*U*).

5.2.2. k

We use three different k: 10, 25, 50

5.2.3. Thresholds

In our application, we would like to be able to classify a new element to several classes at the same time. In order to meet this requirement we introduce two thresholds above which the classification will be validated. The first threshold $t1$ is static. It also makes possible not to assign a class to a new element if the score in the classification rule is too low. The second one $t2$ is a percentage of the best score for each new *LU*. If only this threshold is set, every element is attributed at least to one class.

5.2.4. Semantic spaces

Some semantic spaces are less informative than others due to sparser data or errors in the analysis for specific syntactic relations. Therefore, we apply the classifier with different sets of syntactic spaces: each single representation, all together, all but the *attribute of object*, all but the *attribute of object* and the *attribute of subject* (*all\ atb_obj_subj*), all but the *attribute of object*, the *attribute of subject* and the *adjective complement*, the three best (*direct object, apposition*, and the re-

Resource	Learning data	k	t1	t2	Semantic spaces	Confidence vector	Weighting
EFN.1 *precision*	(U)	10	0	0.95	*all\atb_obj_subj*	D	yes
EFN.2 *coverage*	idem	50	idem	idem	5 best	idem	no

Table 6: Best parameters for enrichment

verse relation of *apposition*), the five best (adding the *noun complement* and its reverse relation).

5.2.5. Confidence vector

We test four distinct ways of computing the confidence vector. The original confidence vector defined by (Kriegel et al., 2005) is:

$$\forall j, 1 \leq j \leq |C| :$$

$$(A)\ cv_{i,j}(o) = \frac{\sum_{u \in sphere_i(o,k) \wedge c(u)=c_j} \frac{1}{d_{norm}(o,u)^2)}}{\sum_{k=1}^{|c|} cv_{i,k}(o)}$$

Paying careful attention to the proposed formula, we notice that the confidence vector used depends on the number of elements in the k-NN sphere that belong to the given class ($|u \in sphere_i(o,k)) \wedge c(u) = c_j|$) but it does not care about the number of elements in each class j. Yet it is more likely that an element belongs to the given class if the class is bigger than the others. Hence we came to the idea of taking the number of elements of the classes into consideration and tested three formulas:

$$(B)\ cvB_{i,j}(o) = \frac{cv_{i,j}}{|c_j|}(o)$$

$$(C)\ cvC_{i,j}(o) = \frac{cv_{i,j}}{log(1+|c_j|)}(o)$$

$$(D)\ cvD_{i,j}(o) = \frac{cv_{i,j}}{log(10+|c_j|)}(o)$$

5.2.6. Weighting

Scores of the translated *LUs* can be used to weight the distances in the classifier. We can either use them or consider all learning data as equivalent.

5.2.7. Setting parameters

As we proceed to an enrichment, we have no gold standard at all to be compared to. In order to set the parameters, we take the global resource (union of the three unfiltered resources) as a comparison resource. We make our parameters vary and compare each new enriched resource by computing the count of *LUs* correctly reassigned over the total count of enriched pairs *LU-frame* present in the comparison resource ($precision$) and how much were reassigned (either correctly or not) over the total number of pairs in the comparison resource ($coverage$). We compute the precision and coverage measure on the comparison resource. We take the resource with the best so-called precision (*EFN.1*) and the one with best so-called coverage (*EFN.2*) to evaluate with the test set.

The best resulting parameters are reported in table 6.

5.3. Results

Evaluation is performed by using as a gold standard the test set of the translation part joined by union with both the chosen enriched resources. We correct the pairs coming from the enriched resource for the test frames. The results are presented in table 7.

	All frames			Test set	
Resource	**#LU-frame**	**#new attributions**	**#Frames**	P	R
Berkeley FrameNet	**11,171**		**796**		
EFN.1 *precision*	9,536	7,581 (79%)[6]	295	82%	7%
EFN.2 *coverage*	27,371	24,997 (91%)[6]	359	61%	10%
TFN + EFN.1[5]	15,132	8,648 (57%)[7]	727	86%	18%

Table 7: Evaluation of the Enriched FrameNets
(5)TFN + EFN.1 = $(Wi\_F_{0.5}max \cap Eu\_F_{0.5}max) \cup EFN.1$
(6) Compared to the comparison resource.
(7) Here are the number and proportion of new attributions after the translation step.

These results appear really good: a precision of 81% while adding 7,581 new *LUs* which is nearly 75% of the size of the Berkeley FrameNet. As regards the resource based on best coverage, we must remind that we process only nouns whereas the translated resource contains any kind of part-of-speech (41% of noun in the Berkeley FrameNet). Our resource $EFN.2$ would have a recall on nouns of nearly 25%. Another reason why recall is not that high compared to the number of existing resource reassigned may be that coverage has been maximized on all frames together while the recall measure of the evaluation is computed on 10

frames separately then averaged.

These enrichments are significant. We can then make the union of one of this resource with a translated one to combine the contributions of the two steps. For example, if we combine our enriched resource *EFN.1* with our translated resource $Wi\_F_{0.5}max \cap Eu\_F_{0.5}max$, we obtain a resource of 15,132 *LU-frame* pairs with a global precision of 86%.

6. Conclusion and future works

We proposed a new approach to transfer the English FrameNet resource to another language and validated this approach for French.

For each dictionary we obtained two types of resources. One is robust (approximately 95% accuracy) but smaller than the Berkeley resource (58% of the number of FrameNet *LU-frame* pairs or 70% if we include FrameNet.fr in our combination). The other one is balanced ($F_{0.5}$ = 67%) and larger than Berkeley FrameNet (34,121 *LU-frame* pairs: almost three times the number of FrameNet *LUs*). Comparison to the existing translation of (Padó and Pitel, 2007) shows that we can produce a resource $Wi \cap EuRADic$ whose size is twice bigger with a better precision (82% instead of 77%).

We also addressed the issue of enrichment to overcome the non-sufficiency of Berkeley FrameNet and of our dictionaries. The enrichment performed on nouns shows really encouraging results and should be tried on verbs which are more often used as predicates.

The results of the evaluation show that our scores reflect well the reliability of an affectation *French LU - Frame*. We can now tackle the annotation task for which we will do a similar translation for the heads of all the role annotated-phrases of the FrameNet corpus. These translations will have a confidence score that will be used by our annotation algorithm.

7. Acknowledgment

We greatly thank Sebastian Padó and Guillaume Pitel for providing us their version of French FrameNet that we use to compare our approach and as a portion of our learning data.

8. References

Collin F. Baker, Charles J. Fillmore, and John B. Lowe. 1998. The berkeley framenet project. In *Proceedings of the COLING-ACL*, pages 86–90.

Romaric Besançon and Gaël de Chalendar. 2005. L'analyseur syntaxique de lima dans la campagne d'évaluation easy. In *Proceedings of TALN 05*.

Ana Fernandez, Gloria Vazquez, Patrick Saint-Dizier, Farah Benamara, and Mouna Kamel. 2002. The VOLEM project: a framework for the construction of advanced multilingual lexicons. In *Proc. of the Language Engineering Conference.*, pages 89–98.

Charles J. Fillmore. 2006. Frame semantics. *Cognitive Linguistics: Basic Readings*, pages 185–238.

Gregory Grefenstette. 2007. Conquering language : Using nlp on a massive scale to build high dimensional language models from the web. In *Proc. of the 8th CICLing Conference*, pages 35–49, Mexico.

Paul Kingsbury and Martha Palmer. 2002. From treebank to propbank. In *Proceedings of the 3rd International Conference on Language Resources and Evaluation (LREC-2002)*, pages 1989–1993.

Hans-Peter Kriegel, Alexey Pryakhin, and Matthias Schubert, 2005. *Database Systems for Advanced Applications*, chapter Multi-represented kNN-Classification for Large Class Sets. Springer Berlin / Heidelberg.

Claire Mouton, Guillaume Pitel, Gaël de Chalendar, and Anne Vilnat. 2009. Unsupervised word sense induction from multiple semantic spaces with locality sensitive hashing. In *Proceedings of RANLP 2009*, Borovets, Bulgaria, September.

Sebastian Padó and Mirella Lapata. 2005. Cross-lingual bootstrapping for semantic lexicons: The case of framenet. In *Proceedings of AAAI-05*, pages 1087–1092, Pittsburgh, PA.

Sebastian Padó and Guillaume Pitel. 2007. Annotation précise du français en sémantique de rôles par projection cross-linguistique. In *Actes de la 14e conférence sur le Traitement Automatique des Langues Naturelles*, page 271.

Marco Pennacchiotti, Diego De Cao, Roberto Basili, Danilo Croce, and Michael Roth. 2008. Automatic induction of FrameNet lexical units. In *Proceedings of Empirical Methods in Natural Language Processing (EMNLP)*, Honolulu, Hawaii, USA.

Benoît Sagot and Darja Fišer. 2008. Combining multiple resources to build reliable wordnets. In *Proceedings of the 11th international conference on Text, Speech and Dialogue*, pages 61 – 68, Brno, Czech Republic.

Karin Kipper Schuler. 2005. VerbNet: A broad-coverage, comprehensive verb lexicon. *Univ. of Pennsylvania-Electronic Dissertations*.