Proceedings of the International FrameNet Workshop 2020: Towards a Global, Multilingual FrameNet, pages 41–47
Language Resources and Evaluation Conference (LREC 2020), Marseille, 11–16 May 2020

FrameNet Annotations Alignment using Attention-based Machine Translation

Gabriel Marzinotto
Orange Labs,
22300 Lannion, France
gabriel.marzinotto @orange.com

Abstract

This paper presents an approach to project FrameNet annotations into other languages using attention-based neural machine translation (NMT) models. The idea is to use a NMT encoder-decoder attention matrix to propose a word-to-word correspondence between the source and the target languages. We combine this word alignment along with a set of simple rules to securely project the FrameNet annotations into the target language. We successfully implemented, evaluated and analyzed this technique on the English-to-French configuration. First, we analyze the obtained corpus quantitatively and qualitatively. Then, we use existing FrameNet corpora to assert the quality of the translation. Finally, we trained a BERT-based FrameNet parser using the projected annotations and compared it to a BERT baseline. Results show modest performance gains in the French language, giving evidence to support that our approach could help to propagate FrameNet data-set on other languages. Moreover, this label projection approach can be extended to other sequence tagging tasks with minor modifications.

Keywords: FrameNet, Machine Translation, Cross-lingual Annotation Transfer, Cross-lingual FrameNet Parsing

1. Introduction

Frame Semantics (Fillmore, 1976) and the FrameNet (Fillmore, 2006) dictionary constitute a valuable resource and a very successful semantic representation scheme, widely adopted and adapted for many NLP applications. For many years a lot of works have studied the adaptability of FrameNet into other languages, showing that many frames are entirely cross-lingual (Gilardi and Baker, 2018). At the same time, FrameNet adaptations for more than 15 languages have been arising. Among these languages, we count Chinese, Danish, Dutch, Finnish, Portuguese, French, German, Hebrew, Hindi, Japanese, Korean, Latvian and Spanish.

Most of these projects use either human translation (*e.g.* Finish (Pedersen et al., 2018) and Danish (Lindén et al., 2017)) while others use semi-automatic alignments (Hayoun and Elhadad, 2016). Also, projects focus on translating the lexical units using bilingual alignments and they tend to deliver a set of annotated examples that is significantly smaller than FrameNet.

More recently, we have seen an initiative to build a multilingual FrameNet lexicon (Gilardi and Baker, 2018), or at least to add relations between similar frames across different languages. Their approach is also based on lexical unit translation using bilingual dictionaries. The objective is to give some guidelines to counter the small divergences we experience today, which are due to the separated evolution of each project.

In this paper, we propose a slightly different example-driven approach to bootstrap FrameNet corpora in new languages using Neural Machine Translation. The main idea is to translate entire annotated sentences instead of lexical units. Then, using neural attention models one can align and project the semantic annotations from the source language (English) into the target language. This allows building a synthetic FrameNet corpus with exemplar sentences. A similar approach have already been studied using Hidden Markov Models (Annesi and Basili, 2010) yielding good results in the English-Italian pair. The advances in NMT allow revisiting this technique using attention models.

Using some post-processing, one can find some of the lexical units that could trigger a frame by looking into the trigger's alignment. Even though this approach is limited to the lexical units for which we have an English sentence example, it allows us to introduce the full sentence in the translation process. This yields a context-aware translation of the lexical units, instead of a one by one word comparison using a dictionary and human experts annotation. We believe both approaches are complementary. This NMT approach can bootstrap a FrameNet lexicon with annotated examples, which can be improved and completed by human experts.

In this paper, we detail the methodology to perform translation and alignment on the English-to-French setting. We give metrics to evaluate this translated corpus and we introduce extrinsic evaluation approach that use the synthetic data-set to train and test automatic FrameNet parsers.

2. Translating and Aligning

Our objective is to automatically produce French translations for both FrameNet and SemEval-07 annotations and provide a methodology that can be extended to other languages for which suitable translation models are available.

A sentence and its translation do not necessarily evoke the same Frames. (Torrent et al., 2018) studies this phenomenon by looking at sentences from the Multi-lingual FrameNet corpus and comparing the frames evoked in the English sentences and the Portuguese translations. They show that, in many cases, the frames evoked on each language differ due to different lexicalization and constructional strategies. Normally, this would imply that FrameNet projections are extremely complex if not unfeasible. We argue that this frame mismatch is widely observed in human translation, but much rare in machine translations (MT). We show this using an example from (Torrent et al., 2018), comparing an English sentence with the Portuguese human translation (HT) and machine translation (MT):

- **EN:** *We have a huge vested interest in it, partly because it's education that's meant to take us into this future that we can't grasp.*
- **PT-HT:** *Nos interessamos tanto por ela em parte porque é da educação o papel de nos conduzir a esse futuro misterioso.*
- **PT-MT:** *Temos um grande interesse, em parte porque é a educação que nos leva a esse futuro que não podemos compreender.*

In Table 1, we list the frames and the lexical units evoked by each sentence. We observe that MT usually does not modify the constructional strategy and is closer to word-by-word translation than the HT sentences. Even though MT sentences may be less sounding, the frames observed in the source and target language tend to be similar. This ensures that the cross-lingual projections can be done, but show that there may be a domain mismatch between natural language and machine translated sentences.

Frame	EN	PT-HT	PT-MT
Size	huge.a	—	grande.a
Stimulus_focus	interest.n	interessar.v	interesse.n
Degree	—	tanto.adv	—
Degree	partly.adv	em parte.adv	em parte.adv
Causation	because.c	porque.c	porque.c
Education	education.n	educação.n	educação.n
Purpose	mean.v	—	—
Performers_roles	—	papel.n	—
Bringing	take.v	conduzir.v	levar.v
Goal	into.prep	—	—
Temporal_colloc	future.n	futuro.n	futuro.n
Certainty	—	misterioso.a	—
Capability	can.v	—	poder.v
Grasp	grasp.v	—	compreender.v

Table 1: Frames and LUs from an English sentence and its Portuguese human (HT) and machine (MT) translations

2.1. Machine Translation Model

To translate the English FrameNet corpus into French we used a state-of-the-art Neural Machine Translation (NMT) algorithm from (Ott et al., 2018) using publicly available pre-trained models from *Fairseq*[1]. This NMT model uses sequence-to-sequence transformers with a total of 222M parameters, it uses a Moses tokenizer, a Word-Piece representation optimized for NMT (Sennrich et al., 2016) and a beam search decoding strategy with a depth of 5. The model achieves state-of-the-art performance on the *newstest2014* test set from *WMT'14* obtaining 43.2 BLEU score on the English-to-French pair.

The most important property of this model is the encoder-decoder attention mechanism (Luong et al., 2015), which is essential for our label alignment strategy. Encoder-decoder attention allows the target-language decoder to look into relevant word-piece information from the source-language encoder. More precisely, for each output word-piece, there is a soft-max distribution vector across the input word-pieces. This distribution shows in which parts of the input the model focus to yield the given output word-piece.

[1]The model is transformer.wmt14.en-fr to be found at `https://pytorch.org/hub/pytorch_fairseq_translation/`

This attention matrix can be used as an indicator of a *soft-alignment* between the word pieces from the input to the output. A simplified example of such an attention matrix is shown in Figure 1. This example is developed in detail in the following subsection.

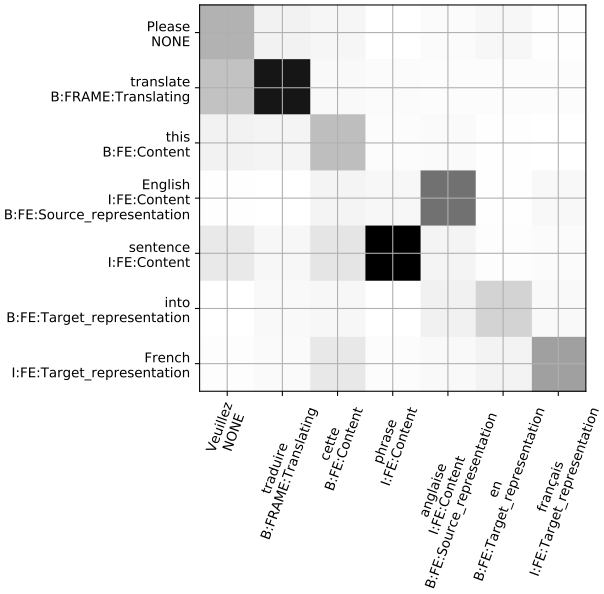

Figure 1: Post Processed Attention Matrix from the NMT

When we look into the raw attention matrix from the NMT we observe that the decoder's attention is distributed between different parts of the input (Ghader and Monz, 2017) and not only on the word-alignment equivalent. For this reason, attention-based alignment is not straightforward. Moreover, in many cases translation inserts tokens that do not have an equivalent. *e.g.* the following sentence and its translation:

"United States helps Australia stop the fire"
"Les États-Unis aident l'Australie à arrêter les incendies"

We observe that the translation adds the definite article for both *"United States"* and *"Australia"* and it also adds a preposition *"à"* to the lexical unit *"stop.v"*. Such word insertions, that are due to some language specific structures, often produce misleading attention vectors. This is the case for the preposition *"à"*, which introduces the goal argument of the verb *"arrêter"*(to stop). Since predicting the word *"à"* depends on both the verb *"aider"* and the role of *"stop"* as the goal, the attention vector for *"à"* is distributed among these two verbs, even though *"à"* is not a viable translation for any of them.

2.2. FrameNet Label Alignment

To generate a translation for the SemEval-07 corpus we translate each sentence using the NMT model described above (2.1.). For each sentence, we recover the attention matrix and apply the following post-processing steps:

Re-establish tokenization: Since the NMT model uses word-piece representations, the first step is to project the attention matrix into a full-word form. To do this, we perform sub-matrix sums on the sets of rows (and columns)

that correspond to the same input (and output) words.

Part-of-Speech (POS) weighting: To avoid alignment mismatches as those described above (2.1.), we perform POS and dependency parsing on both the input and the output sentences using spaCy[2]. SpaCy has close to state-of-the-art performance on these tasks. We use the POS to post-process the attention matrix applying two simple rules. First, we mask structural POS such as `"PUNCT"`, `"DET"`, `"INTJ"`, `"SYM"`, `"X"`, `"AUX"`, `"PART"`. We do this because the attention vectors for these POS are hard to align due to their structural nature (see 2.2.). Second, we encourage the alignment between words with the same POS. To do this, we multiply by 10 the attention matrix entry (i, j) if the input and output words $(w_i\ w_j)$ have the same POS. Finally, we normalize the matrix by columns, *e.g.* we divide each entry (i, j) by the sum of the values in column j. This allows interpreting each matrix entry as the percentage of attention the word w_j pays to the input w_i.

Label Projection: To project annotations from the English FrameNet into the French translation we flattened the FrameNet annotations over the English word tokens. Then, we paired each output word to the input word with the highest attention score and propagated the input labels into the matching output token, as shown in Figure 1. Since the model is not equally confident in every translation, we score each label projection with the value of the attention coefficient between the input and the output word.

Confidence Threshold: To decide which labels project and which reject, we apply a threshold on the confidence score. The choice of the threshold is not straightforward. If it is too low, it introduces alignment mismatches, which can be seen as frame and frame element insertions. On the other hand, a high value will only project a few annotations. We chose the threshold that maximizes the harmonic mean between the the number of annotations projected and the amount of duplicated projections. This step is explained with more detail in Section 2.3..

frame element Completing: To ensure homogeneity in the spans of the frame elements, we used the syntactic dependency parsing to complete the spans of the frame elements applying two simple rules. First, if a determinant or a preposition is attached to a frame element through its syntactic parent, it inherits the label of that parent. Second, for words masked during the POS weighing process (*e.g.* for being either `"PUNCT"`, `"DET"`, `"INTJ"`, `"SYM"`, `"X"`, `"AUX"` or `"PART"`), we assign them a frame element label if words that precede and follow are part of the same frame element. This allows us to merge potential frame element segments, that got split during the alignments.

This sequence of steps is language independent and fairly easy to implement. However, this does not mean the corpora translation process is flawless. In the next sub-section we study the quality of the generated corpus.

2.3. Generated Corpus Analysis

In this subsection, we explore the translated corpus and establish some comparison with the original corpus.

[2]spaCy website: `https://spacy.io/`

First, we evaluate the projection using deletion and insertion metrics. To do so, we assign an id to each annotation (frame or frame element) of the SemEval-07 corpus. Then, we use these ids to evaluate the French translation by counting how many ids were lost (deletions) and how many ids got duplicated (insertions). This is a rough metric, somewhat similar to standard precision/recall evaluations. However, it does not imply that annotations are aligned to the right words in the target language. We only measure if the algorithm finds suitable candidates to project the annotations. We use this metric because more precise evaluations require skilled human annotators/validators. This evaluation is strongly tied to the confidence threshold selected during the alignment (see 2.2.). In Figure 2 we observe the deletion/insertion trade-off of the projection computed using different values of the confidence threshold. This trade-off achieves a maximal F1 score of 85.5%. At this point, the insertion metric is 90.1%, meaning that there are no more than 10% of false insertions, while the deletion metric is 81.4%. meaning that we project about 81% of the annotations and delete 19%. These alignment performances show that there is still room for improvement, either via better NMT models or through more complex post-processing strategies. However, in this paper, we settled at this 85.5% F1 and we evaluate how useful this simple approach can be.

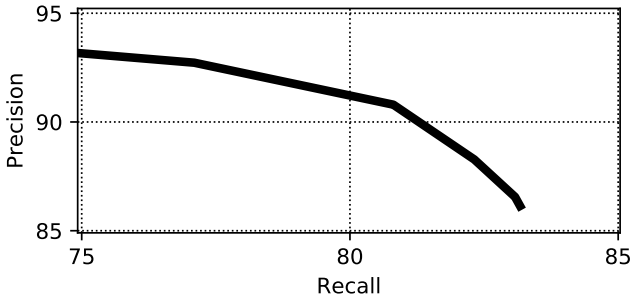

Figure 2: Insertion Deletion Trade-off of the label projection

After this evaluation, we look into more detail which parts of the projection were problematic. In Figure 3, we show the distribution of frames that introduced the highest amount of errors during the translation. Whenever the French column is larger(smaller) that the English column it means that the frame got inserted(deleted) several times.

We observe that frames such as `Existence`, `Quantity`, `Attributed_information`, `Capability` and `Partitive` suffer several deletions. Many times, these deletions are due to alignment constrains (see 2.2.) such as not projecting labels from auxiliary verbs. *e.g.* The frame `Existence` in the expression *"There was a time..."* translates into *"Il fut un temps..."*, here the auxiliary verb *"fut"* was masked. Also, for the frame `Quantity` in the sentence *" Brazil helped several countries..."* which translates into *"Le Bresil a aidé plusieurs pays ..."* the word *"plusieurs"* is a determinant which gets masked. Some of these errors could be fixed by introducing more language specific projection rules. As for the others, changes in the constructional strategies make label projection very difficult and error-prone.

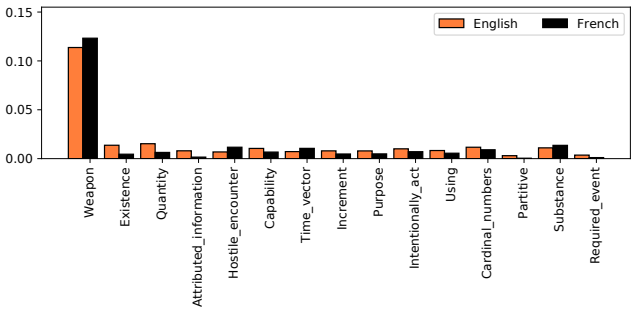

Figure 3: Distribution of the frames with the highest number of projection errors. Whenever the French column is larger(smaller) than the English column it means that the Frame got inserted(deleted) several times.

On the other hand, frames such as `Weapon`, `Hostile_encounter`, `Time_vector` and `Substance` suffered insertions. These insertions were mostly due to lexical units with prepositional or adverbial POS. As we discussed previously, the attention vectors for these words are hard to align and in some cases, they get projected twice.

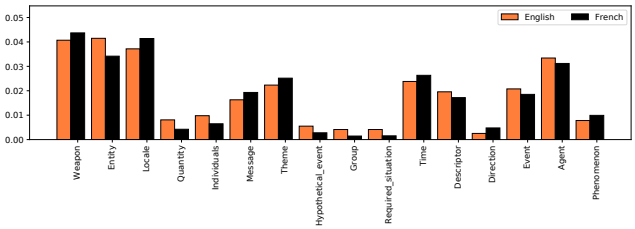

Figure 4: Distribution of the frame elements with the highest number of projection errors

When we repeat the same analysis on the frame element translation we observe that many frame elements projection errors are the proliferation of a frame error. This is expected since we filter frame elements without a trigger. There are a few exceptions for this rule, such as time, agent and theme, which are quite generic frame elements. They can get deleted due to translation errors or due to low confidence scores. For an example of a translation error, consider the sentence *"... due to the urgency, its development was fast ..."* that gets translated into *"... face à l'urgence, le development a été rapide..."*. Even though it is an acceptable translation, in the English sentence, the word *"its"* is an agent frame element. When we look into the translation, there is no match for this word (it was deleted). Maybe, a more accurate translation would add a pronoun *"leur"* as in *"... face à l'urgence, leur development a été rapide..."*. This sort of error does not seem dangerous, as the final sentence is still fully annotated.

Beam Search: an interesting aspect of this approach is that we can use NMT along with beam search decoding to extract several translation candidates for any sentence. This is particularly useful when building FrameNet dictionaries, as it allows us to generate translations with different lexical triggers taking into account the sentence's context. On the other hand, this technique can also be used to do data augmentation on the target language. We did not explore this option in our paper, and we have left it for future work.

3. `biGRU+BERT` Semantic Parser

We propose a `biGRU+BERT` model architecture, inspired from state-of-the-art models in Semantic Role Labeling (He et al., 2017) and FrameNet parsing (Yang and Mitchell, 2017; Marzinotto et al., 2018b; Marzinotto et al., 2019). Our architecture, uses 2 layers of bidirectional `GRU` stacked on top of a pre-trained *'bert-base-multilingual-cased'* model from Huggingface[3]. A diagram of our model architecture is shown in figure 5. To encode semantic labels into flat structures we use a `BIO` encoding scheme. To ensure that output sequences respect these `BIO` constrains we implement an A* decoding strategy similar to the one proposed by (He et al., 2017).

To deal with sentences containing multiple lexical units we have built training samples containing only one trigger. More precisely a sentence containing N triggers provides N training samples. The downside of this approach is that during prediction time, parsing a sentence with N triggers requires N model applications. At decoding time every pair of { sentence, trigger } is processed by the network to output a probability distribution on the frames and FE for each word. Then, we apply a *coherence filter* to these probabilities to make sure that the predicted frame elements are compatible with the predicted frame by filtering the extraneous frame elements. This *coherence filter* chose the frame with the highest probability on the trigger and uses it as the predicted frame (represented as the label assigned by the tagger to the trigger). Then, given that frame, the *coherence filter* masks all the frame element labels that are incompatible with the selected Frame.

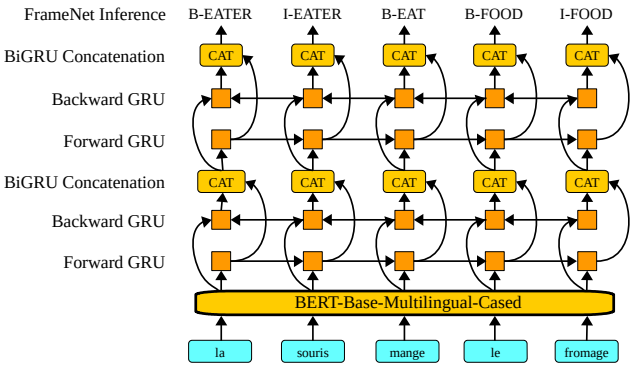

Figure 5: `biGRU` Model Diagram

4. FrameNet Parsing Experiments

4.1. Data

In our experiments we use 4 FrameNet corpora:

`SemEval07-EN`: A corpus of full-text annotations from FrameNet project used for the shared task 19 from `SemEval-07` (Baker et al., 2007). This corpus consists of annotated journals and it contains 720 different Frames.

[3] `https://huggingface.co/transformers/`

	Lang	Data Source	Nb. Diff. Frames	Nb. Diff. LU	Nb. Diff. FE	Nb. Diff. Words	Nb. Sentences w/LU	Nb. Annot. LU	Avg. Nb. Annot. per LU
SemEval-07	EN	Journals	720	3,197	754	14,150	4,020	24,770	7.7
ASFALDA	FR	Journals	121	782	140	33,955	13,154	16,167	20.7
CALOR	FR	Encyclopedias	53	145	148	72,127	22,603	31,440	215.4

Table 2: Summary and statistics about the existing FrameNet corpora. Columns from left to right:
(1) Language. (2) Source of the non-annotated data. (3) Number of different frames annotated at least one time.
(4) Number of different LU annotated at least one time. (5) Number of different FE annotated at least one time.
(6) Number of sentences containing at least one annotated LU. (7) Total number of LU instances annotated.
(8) Average number of annotations per LU.

SemEval07-FR: This corpus is the translation of SemEval07-EN into French following the methodology described in Section 2.2.

ASFALDA French FrameNet: is the first French FrameNet project (Djemaa et al., 2016) which outlines and produces a FrameNet equivalent for the French language. This corpus gathers experts frame annotations on sections of the journal *Le Monde*, it contains 121 different frames that focus on four notional domains: commercial transactions, cognitive positions, causality, and verbal communication. ASFALDA tries as much as possible to align with the English FrameNet structure; however, it also introduces new frames whenever the differences in both languages do not allow conciliation.

CALOR: is a publicly available corpus(Marzinotto et al., 2018a) of French encyclopedic documents human-annotated using FrameNet semantics. This corpus was designed from the perspective of Information Extraction tasks. Like ASFALDA, CALOR uses a *partial parsing* policy, in which annotations are limited to a small subset of frames from FrameNet. The CALOR corpus contains 53 different frames selected as the most representative and frequent within the annotated documents. Despite the small number of Frames, CALOR is the corpus with the largest set of annotated lexical units.
Table 2 summarizes relevant statistics on each corpus.

4.2. Evaluation setting

We run experiments using the standard Train, Validation, and Test of each corpus. For the SemEval07-FR corpus, these subsets are equivalent to the ones from SemEval07-EN. To evaluate our models we use 2 of the main sub-tasks of FrameNet parsing:

- **Frame Identification:** Consists in selecting the frames evoked by each of the lexical units in the sentence. One frame per lexical unit. Here, we use the gold annotated lexical units, and we do not try to infer them from raw text.
- **Argument Identification:** Consists in finding the spans of words that correspond to semantic roles and assigning them the correct frame element labels from the selected Frame.

Since the set of lexical units on each corpus is different, and since this difference is due to arbitrary choices about what should be annotated in the *partial annotations* schemes from CALOR and ASFALDA we consider that all the lexical unit instances are known and given as input to our model. For the same reasons, we discard lexical unit annotations that refer to the frame OTHER, which is an artifact to handle frames out of the scope of the *partial annotation*. In this setting, each lexical unit triggers a frame from the frame dictionary of the given corpus.
We score our models using:

- **Frame Identification:** Accuracy on the frame classification Task with gold lexical units. FrameNet official evaluation scripts use the frame hierarchy to introduce a matching metric that gives partial credit when predicting related frames (*e.g.*, a more generic Frame). Since this hierarchy is no available on each corpus, we evaluate using exact frame matching and we do not exploit any of the frame-frame relations proposed in FrameNet.
- **Argument Identification:** Precision, Recall, and F1 or the frame element detection. This metric can be computed either using the gold or the predicted Frames. In the official evaluation scripts, the token span of the hypothesis must be the same as in the reference for a frame element to be correct (Hard Spans or H-Spans). We have loosened this constraint to introduce a new variant of the evaluation metrics. Instead of demanding exact span match, we use a weighted correctness score proportional to the overlap ratio between the gold (S_{ref}) and predicted (S_{hyp}) spans (W-Spans) computed using equation 1.

$$W_{span}(S_{ref}, S_{hyp}) = \frac{|S_{ref} \cap S_{hyp}|}{|S_{ref} \cup S_{hyp}|} \quad (1)$$

5. Bi-lingual Semantic Parsing Experiments

Recent works (Pires et al., 2019) have shown that BERT language models pre-trained on multilingual corpora have a fairly good performance in a zero-shot cross-lingual transfer setting. More precisely, if we fine-tune a multi-lingual BERT using task-specific annotations from a monolingual corpus and then evaluate the model in a different language, the system will be able to generalize to the new language, up to some extent.

	Test: ASFALDA		Test: CALOR	
	Frame Id.	Arg Id. pred frames	Frame Id.	Arg Id. pred frames
TRAINING CORPUS	ACC	F1-W	ACC	F1-W
CALOR	-	-	98.8	67.8
ASFALDA	73.3	52.6	-	-
SemEval-07-EN	42.2	**19.7**	77.5	30.5
SemEval-07-FR	**73.3**	19.1	**78.1**	**31.0**
SemEval-07-EN+FR	**74.8**	**20.9**	**79.4**	**32.4**

Table 3: Performance of a biGRU+BERT on the Gold French FrameNet corpora using different training data-sets

In this experiment, we harness this zero-shot generalization property to evaluate the quality and the relevance of our corpus translation strategy extrinsically.

	Frame Id.	Argument Id. w/pred frames		
	ACC	P	R	F1
SemEval-07-EN	85.3	49.6	40.3	44.5
SemEval-07-FR	80.8	29.8	15.4	20.3
CALOR	77.5	37.3	25.7	30.5
ASFALDA	42.2	24.5	16.4	19.7

Table 4: Performance of a biGRU+BERT model trained on the SemEval-07-EN corpus and tested on other corpora

	Frame Id.	Argument Id. w/pred frames		
	ACC	P	R	F1
SemEval-07-FR	80.5	38.5	26.0	31.1
SemEval-07-EN	82.8	37.3	23.3	28.7
CALOR	78.1	37.7	26.3	31.0
ASFALDA	73.3	38.6	15.8	19.1

Table 5: Performance of a biGRU+BERT model trained on the SemEval-07-FR projected corpus

First, we train a model on the SemEval-07-EN corpus and test in on all the available corpora to establish a zero-shot baseline performance. We train the model for 40 epochs and we used two validation sets, one from SemEval-07-EN and the other form SemEval-07-FR. This way, we retained the best performing model for each validation set and language. We observed that the validation error stops decreasing earlier on the French corpus than on the English one. This could be expected due to the monolingual training configuration. However, we did not observe significant over-fitting or catastrophic forgetting on the French language when doing supplementary iterations. After training, we used the best English model to produce inferences on the SemEval-07-EN test and the best French model to produce inferences on the 3 French corpora. We evaluate using precision, recall and F-score on Weighted Spans (see Section 4.2.). The results for this experiment are shown in Table 4. We observe that transfer learning and the language similarities between English and French can bootstrap a low-performance baseline for French. The system is surprisingly good at detecting Frames. However, it has very low performance on the full FrameNet parsing Task, showing a particularly low recall.

In the following step, we train a model using the projected SemEval-07-FR corpus. The performance of this model is given in Table 5. We observe that the French model trained on the synthetic data-set achieves slightly better performance than the English model baseline. It is better at the Frame Identification step, and it is slightly better in terms of Argument Identification as well. Moreover, the French model is capable of generalizing back to the English language, showing close performances between SemEval-07-EN and SemEval-07-FR. The fact that it is easier to generalize toward English can be interpreted in several ways. One possibility is that the French FrameNet parsing task is more difficult than its English equivalent since FrameNet was designed for English or because French is a more verbose language. Another possibility is that due to the alignment errors, the SemEval-07-FR test data-set flaws would be penalizing good predictions.

In Table 3 we present the scores for the hand-annotated French data-set for different configurations varying the training corpus. We observe that even though the French model trained on the synthetic data-set achieves slightly better performances than the English baseline, we are still far from the state-of-the-art performances on each corpus obtained through training a FrameNet parser on hand-annotated French data-sets. However, part of this performance gap is due to the differences in the number of Frames. A SemEval-07 model handles 720 different frames and 3197 different lexical units, therefore it is much more prone to choosing a wrong frame than a CALOR model, which handles only 53 Frames. Another reason for this performance gap may be the domain changes, CALOR and ASFALDA contain natural language sentences instead of translated sentences. Moreover, previous experiments on the CALOR corpus from (Marzinotto et al., 2019) have shown that even within the same data source, domain changes yield around 10% F1 performance drop on the Argument Identification Task. Finally, we trained a model combining both SemEval-07-EN and SemEval-07-FR and obtained small gains on the French corpora, showing that the translated data adds some supplementary information to the model.

6. Conclusion

This paper presented a simple method to project FrameNet annotations into other languages using attention-based neu-

ral machine translation (NMT) models. We tested our approach on the English-to-French configuration showing that 90% of the labels can be easily projected without introducing much noise. We performed an in-depth analysis of the French corpus obtained through translation and we showed the most common projection errors. Then, we use existing French FrameNet corpora to assert the quality of the translation. We trained a BERT-based FrameNet parser using the projected annotations and compared it to a BERT baseline showing modest gains on French. All these results support that our approach could help to propagate FrameNet dataset on other languages where sufficiently developed NMT models exist. Moreover, this label projection approach can be extended to other sequence tagging tasks with minor modifications.

7. Bibliographical References

Annesi, P. and Basili, R. (2010). Cross-lingual alignment of framenet annotations through hidden markov models. In Alexander Gelbukh, editor, *Computational Linguistics and Intelligent Text Processing*, pages 12–25, Berlin, Heidelberg. Springer Berlin Heidelberg.

Baker, C., Ellsworth, M., and Erk, K. (2007). Semeval'07 task 19: Frame semantic structure extraction. In *Proceedings of the 4th International Workshop on Semantic Evaluations*, SemEval '07, pages 99–104, Stroudsburg, PA, USA. Association for Computational Linguistics.

Djemaa, M., Candito, M., Muller, P., and Vieu, L. (2016). Corpus annotation within the french framenet: methodology and results. *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC)*, X:3794–3801.

Fillmore, C. J. e. a. (1976). Frame semantics and the nature of language. In *Annals of the New York Academy of Sciences: Conference on the origin and development of language and speech*, pages 20–32.

Fillmore, C. J. e. a. (2006). Frame semantics. *Cognitive linguistics: Basic readings*, 34:373–400.

Ghader, H. and Monz, C. (2017). What does attention in neural machine translation pay attention to? *CoRR*, abs/1710.03348.

Gilardi, L. and Baker, C. (2018). Learning to align across languages: Toward multilingual framenet. In *Proceedings of the International FrameNet Workshop*, pages 13–22.

Hayoun, A. and Elhadad, M. (2016). The hebrew framenet project. In *LREC*.

He, L., Lee, K., Lewis, M., and Zettlemoyer, L. (2017). Deep semantic role labeling: What works and what's next. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*.

Lindén, K., Haltia, H., Luukkonen, J., Laine, A. O., Roivainen, H., and Väisänen, N. (2017). Finnfn 1.0: The finnish frame semantic database. *Nordic Journal of Linguistics*, 40(3):287–311.

Luong, M.-T., Pham, H., and Manning, C. D. (2015). Effective approaches to attention-based neural machine translation. *arXiv preprint arXiv:1508.04025*.

Marzinotto, G., Auguste, J., Bechet, F., Damnati, G., and Nasr, A. (2018a). Semantic Frame Parsing for Information Extraction : the CALOR corpus. In *LREC2018*, Miyazaki, Japan, May.

Marzinotto, G., Béchet, F., Damnati, G., and Nasr, A. (2018b). Sources of Complexity in Semantic Frame Parsing for Information Extraction. In *International FrameNet Workshop 2018*, Miyazaki, Japan, May.

Marzinotto, G., Damnati, G., Béchet, F., and Favre, B. (2019). Robust semantic parsing with adversarial learning for domain generalization. In *Proc. of NAACL*.

Ott, M., Edunov, S., Grangier, D., and Auli, M. (2018). Scaling neural machine translation. *CoRR*, abs/1806.00187.

Pedersen, B. S., Nimb, S., Søgaard, A., Hartmann, M., and Olsen, S. (2018). A danish framenet lexicon and an annotated corpus used for training and evaluating a semantic frame classifier. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Pires, T., Schlinger, E., and Garrette, D. (2019). How multilingual is multilingual bert? *CoRR*, abs/1906.01502.

Sennrich, R., Haddow, B., and Birch, A. (2016). Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany, August. Association for Computational Linguistics.

Torrent, T. T., Ellsworth, M., Baker, C., and Matos, E. (2018). The multilingual framenet shared annotation task: a preliminary report. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation*, pages 62–68.

Yang, B. and Mitchell, T. (2017). A joint sequential and relational model for frame-semantic parsing. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1247–1256. Association for Computational Linguistics.