Explainable Projection: Cross-lingual semantic role labeling

Anonymous ACL submission

Abstract

Semantic role labeling (SRL) is a central task in many applications, e.g., machine translation, question answering, summarization, and more recently, complex tasks such as stance detection. However, cross-lingual projection of SRL labels has remained a thorny problem in NLP. The scarcity of semantically annotated corpora makes it difficult to build semantic role labelers, particularly for languages where hand-annotated labels are not readily available. We leverage semantic isomorphism at the level of predicate-argument structure to induce SRL systems from unlabeled bilingual corpora. We demonstrate that this approach yields explainable representations that readily project to new languages. Our novel contribution is the use of a simple word-to-word alignment followed by a *First Come First Assign* (FCFA) algorithm and a handful of linguistically-informed constraints specified at the predicate-argument level. These constraints provide a systematic mapping to semantic-role *divergence* categories that serve as the basis for analysis of our FCFA approach. A two-step process rapidly produces explainable SRL output: simple alignment followed by application of FCFA. This approach yields SRL projection results that are comparable to state of the art performance (XSRL), but without relying on complex transformer-based scoring schemes for multi-word alignments.

1 Introduction

Semantic role labeling (SRL) is a high level natural language processing (NLP) task that captures "who did what to whom". SRL labels are used in downstream tasks such as machine translation, question answering, summarization (Liu and Gildea, 2010; Genest and Lapalme, 2011), or more complex tasks such as extraction of privately held *beliefs* and *stances* (Mather et al., 2021, 2022).

SRL has been extensively studied in English due to the high availability of English-specific SRL annotated datasets (Fei et al., 2020a). However, scarcity of SRL-annotated corpora in other languages motivates the need for cross-lingual approaches that project SRL labels from English to other languages, to produce non-English SRL-labeled corpora. Existing methods of SRL transfer from English to a target language use model transfer (Kozhevnikov et al., 2013; Fei et al., 2020b), syntactic rules (He et al., 2019; Li et al., 2018), and projection (Padó et al., 2009; Fei et al., 2020a).

Two projection methods are generally employed: (a) translation-based and (b) alignment-based. Translation-based translates English data into the target language and transfers the English labels. This approach has shown good performance due to recent improvements in neural machine translation (NMT) models (Fei et al., 2020a; Gehring et al., 2017; Hassan et al., 2018).

Unfortunately, NMT is not always optimal for building cross-lingual SRL corpora due to translation *divergences* between source and target languages (Dorr, 1994; Lee et al., 2018), which affect SRL performance. Moreover, translation-based projection generally involves tree-to-tree mappings to build cross-lingual SRL-annotated corpora (Prazák et al., 2017), yet these mappings are hampered by a lack of isomorphism at the syntax level (Shen et al., 2016), and they are often more trouble than they are worth, given that SRL does not require the full power of tree/graph-based representations. Finally, state-of-the-art projection approaches such as XSRL (Daza and Frank, 2020) employ a lightweight and portable word-to-word alignment, but without a framework for explaining output.

Our work uses word-to-word alignment-based projection from English to French, with an eye toward addressing the issues above. The approach enables the transfer of SRL labels from source to target sentences, and creates a French SRL-annotated corpus. For the purpose of this paper, we include the following SRL labels: ARG0 (Agent), ARG1

(Patient), ARG2 (beneficiary or instrument), ARG3 (start point or attribute), and also VERB (main predicate). Accurate SRL projection requires resolution of translation divergences. State-of-the-art XSRL classifies and addresses two divergences types: Nominalizations and Separable Verb Prefixes—a subset of those explored in this paper.

Existing studies introduce unnecessary complexity (which impacts speed) or require human labor to address divergences, e.g., syntactic constituents extraction (He et al., 2019; Padó et al., 2009), probability distributions application (Akbik et al., 2015), BERT-score calculation(Zhang et al., 2019), or human intervention (Daza and Frank, 2020). Instead, our approach defines and addresses a general set of predicate argument divergences, following those described by Dorr (1994).

(a) Categorial: ***is enough*** - ***suffit***
[V-is] --------> [V-suffit]
[ARG2-enough] --------> [V-suffit]

(b) Light Verb: ***trust*** - ***fait confiance***
[ARG0-we] --------> [ARG0-on]
[V-trust] --------> [V-confiance]

(c) Structural: ***prevent*** - ***évitera à la***
[V-prevent] --------> [V-éviter]
[ARG1-France] --------> [ARG1-France]

Figure 1: Divergence types

Fig. 1 illustrates three representative examples of generalized divergences: (a) categorial, (b) Light Verb, and (c) Structural. categorial divergences are mappings where the source and target have different parts of speech (POS). In (a), the adjective *enough* (coupled with the verb *is*) maps to the verb *suffit* (i.e., *suffice*). During SRL processing, the word *is* is assigned VERB and *enough* is assigned ARG2, on the French side, *suffit* is appropriately assigned VERB with no ARG2. In (b), *trust* is mapped to *fait confiance* (i.e., *have trust*). Here, *fait* is "light" in meaning, with the main semantic content conveyed by *confiance*. Our approach assigns VERB to *confiance*, leaving *fait* appropriately unassigned. Structural divergence is illustrated in Fig. 1 (c), where an incorporated argument (*France*) appears in the target language as the object of a preposition *a la France*. Here, *prevent* maps to *éviter à*. Our approach assigns VERB only to *éviter*, leaving *à* appropriately unassigned.

Our process does not require transformer-based scoring or syntactic rules to address these label divergences. Instead we apply a simple, efficient algorithm that retains accuracy and induces explainability. We use word-to-word alignment along with a queued rule-based SRL projection model: *First Come, First Assign* (FCFA). This approach avoids noise introduced by divergent mappings. We take advantage of isomorphic properties at the semantic level, using predicate-argument structures to systematically map potentially divergent semantic-role labels, while retaining underlying predicate-argument representations (often more than one) to support explainability in reporting results.

Additionally, we provide a visualization tool that displays these linguistically-motivated representations, one for each predicate-SRL labels. These visualizations display how and why each SRL label assignment was made and reveals errors for easy remediation. We create a SRL-annotated dataset and build a French semantic role labeler without reliance on a gold-standard human-annotated corpus. Our results show the FCFA model (a) performs 107% faster than XSRL, (b) offers F1 score comparable to XSRL, and (c) is more explainable, via our visualization tool.

Next we discuss background for our work. Section 3 provides our methodology for explainable projection followed by experiments and analyses in Section 4. We present limitations in Section 5, followed by conclusions and future work in Section 6, and end with ethical considerations.

2 Background

We highlight three background areas: cross-lingual semantic role labeling, pre-training of SRL models, and explainability in NLP.

2.1 Cross-lingual Semantic Role Labeling

Cross-lingual semantic role labeling is generally achieved through model transferring **or** annotation projection. Model transferring approaches do not attempt to align datasets across languages. Hence, these approaches involve the creation of a separate dataset for each language. McDonald et al. (2013) generate syntactic-dependency datasets for six languages. Transfer models use shared feature representations (Kozhevnikov et al., 2013). Polyglot SRL (Mulcaire et al., 2018) employs word vectors and is trained on annotation union between two languages. One cross-lingual encoder-decoder model

translates and assigns SRL at the same time for resource-poor languages (Daza and Frank, 2019).

Projection-based methods are used for tasks such as POS tagging (Yarowsky and Ngai, 2001), dependency parsing (Hwa et al., 2005) and machine translation (Zhang et al., 2008; Shen et al., 2016). Tree-to-tree mapping (Shen et al., 2016) is used in machine translation and projects dependency trees from source to target without regard to isomorphic predicate-argument structures. Tree sequence alignment takes advantage of syntax-based and phrase-based methods (Zhang et al., 2008). Projection of universal dependency trees along with cross-lingual features have been used in Prazák et al. (2017) to develop an SRL system.

SRL Projection can be done without relying on tree-to-tree mappings as semantic information abstracts away from surface structure and thus is less prone to syntactic variations (Padó et al., 2009). Word-to-word alignment-based projection demonstrates degrees of success (Daza and Frank, 2020; Fei et al., 2020a; He et al., 2019). However, such approaches introduce considerable noise due to translation divergences and alignment errors (Akbik et al., 2015). Additional resources such as syntactic constituents (Padó et al., 2009) and projection probability distributions (Akbik et al., 2015) are used to improve the performance of word-alignment based projection annotations. Noise is further reduced by using gold-standard annotated source data and projecting the relevant annotations to translated target sentences. Translation-based projection has proven useful (Fei et al., 2020a).

Our French SRL labeler relies on central elements of projection-based models. When the projection-based model aligns two sentences in different languages, one-to-one word mappings are not guaranteed. XSRL (Daza and Frank, 2020) classifies many-to-one mappings into three cases: nominalizations, light verb construction, and separable verb prefixes. For test data, XSRL uses human validation to align two languages. They then use BERT-score (Zhang et al., 2019) to automatically project SRL labels from source to target via score-based voting to transfer labels. Björkelund et al. (2009) build an ML classifier using logistic regression to implement multilingual semantic role labeling. Syntactic rules are also used to build multilingual SRL models (He et al., 2019).

While our approach also relies on projection, it differs from above as it explains three divergent cases and provides visualization to illuminate our process for SRL transfer. We then compare performance of our French semantic role labeler with that of XSRL, a state-of-the-art implementation and community standard for projection-based multilingual SRL model.

2.2 Pretrained SRL model

Deep learning has recently been explored for SRL. For example, He et al. (2017) use a BiLSTM and achieve high performance. Deep neural network models have mainly explored the English SRL task, yet challenges still remain for languages where hand-annotated labels are not readily available.

To lessen this gap, Mehta et al. (2018) implement a semi-supervised SRL model using syntactic parse trees and BERT-based models trained for SRL (Shi and Lin, 2019). Their motivation for using BERT-based SRL is that BERT leverages the capabilities of a pre-trained language model rather than using additional lexical and syntactic data such as POS tags (Marcheggiani et al., 2017) and syntactic tree structures (Roth and Lapata, 2016; Li et al., 2018).

In addition to demonstrating the effectiveness of pre-trained language models for SRL, Shi and Lin (2019) also put forth a unified representation for argument annotation by combining both span-based and dependency-based annotation schemes.

Works described above have improved the performance of SRL, however, they are mostly implemented in English. Our approach uses BERT-based alignment to project French SRL labels from English and provide French corpora that are used as training data for a French SRL model.

2.3 Explainability in NLP

Explainable NLP has attracted significant attention in the NLP community (Søgaard, 2021). Since deep learning or embedding-based approaches began, many NLP research efforts have fallen under the umbrella of *black box* models, whereas traditional NLP techniques, such as decision tress, rules, hidden Markov models etc. are recognized as *white box* techniques due to their inherent explainability.

Danilevsky et al. (2020) argue that methods used to generate or visualize explanations are significant ways to characterize explanations, and they present various ways to represent explainability such as raw examples, declarative rules, or saliency highlighting.

Explainability has been studied for various tasks and domains. Liu et al. (2020) suggest a genera-

tive explanation framework for classification, with fine-grained explanations. Healthcare implements explainability by classifying clinical data and visualizing methods and model-agnostic post-hoc attributions (Danilevsky et al., 2021). Nevertheless, explainability for semantic role labeling is scarce. Our visualization tool lessens this gap by providing representations of the SRL transfer process.

3 Methodology for Explainable Projection

We introduce an explainable projection approach that uses word-to-word alignment coupled with *First Come First Assign* (FCFA). Our approach supports explainability in that the decisions behind label projections are clearly displayed rather than hidden behind *black box* algorithms. We use CoNLL-U format introduced by More et al. (2018), including ten fields, such as word segmentation, POS tagging, and morphological features, etc., and we assign SRL labels to the eleventh column.[1] Fig. 2 depicts intermediate predicate-argument representations. A predicate appears on top of each representation. Our projection technique is shown using yellow (English) on the left and blue (French) on the right as shown at each representation.

To achieve explainable projection, we adopt a framework that creates SRL annotations for the French non-gold standard datasets as follows. We translate French data to English using Google Translate API (Han, 2015). The English translated data is then assigned SRL labels using the SRL predictions of AllenNLP SRL model (SRL-BERT) (Shi and Lin, 2019). FCFA then projects the English SRL labels to the French dataset.

The French dataset is then used as training data. We use AllenNLP's trainer, with SRL-BERT (Shi and Lin, 2019) as the model algorithm. Performance of FCFA projection is compared to state-of-the-art XSRL projection and demonstrates that FCFA projection is two times faster than XSRL projection while remaining equally as accurate.

Analysis of FCFA and XSRL outputs uses precision, recall, and F1 score. FCFA projection achieves F1 scores comparable to the XSRL (50.8 Twitter, 59.6 Wikipedia), while retaining speed gains and adding explainability.

Two products of our implementation are: (a) a SRL labeled bilingual corpus (French-English as our test case) to train a French SRL model; and (b) a set of linguistic representations (one for each predicate-role assignment) that provides a window into why/how the system produces its output, while elucidating errors that can be readily remedied.

3.1 Data for Experimental Design

Previous projection-based SRL models focus on pre-annotated SRL data in news genre. For example, CONLL-09 and Proposition Bank (Daza and Frank, 2020; He et al., 2019; Fei et al., 2020a).

Our work expands to social media via French Twitter data (Daignan, 2017).[2] Emojis, URLs, and mentions (@id) are removed. Wikipedia data, collected using Selenium (García et al., 2020) is also used.[3] Wikipedia data focuses on French Elections and politics, particularly around the Russia-Ukraine war. Both datasets are preprocessed via spaCy[4] and transformed into CoNLL format (Buchholz and Marsi, 2006) readying them for projection.

3.2 Source (English) corpora

Projection requires source corpora that are already annotated with SRL labels, typically gold standard annotations. However, our source data do not contain any SRL labels let alone gold standard labels. Therefore, we need to assign SRL labels to our source data to ready them for projection.

To create a source corpora annotated with SRL labels, first, we translate French datasets into English using googletrans 3.0.0 API (Han, 2015),[5] which is a free and unlimited python library for Google Translate API. We use SRL-BERT (Shi and Lin, 2019)[6] from AllenNLP to assign SRL labels to the English corpora, and then these SRL labels are projected to the French corpora. AllenNLP's SRL-BERT model achieves an F1 Score of 86.49 on the English Ontonotes dataset (Shi and Lin, 2019).

3.3 Word-to-word Alignment

We implement word-to-word alignment (Bacciu, 2021) to project SRL labels. This algorithm is

[1] CoNLL-U format described at https://universaldependencies.org/format.html

[2] https://www.kaggle.com/datasets/jeanmidev/french-presidential-election, data include user names, but is open and publicly available.

[3] Selenium Apache License 2.0 https://www.selenium.dev/documentation/about/copyright/#license

[4] spaCy 3.4.1 with en_core_web_md and fr_core_news_md is used for tokenization

[5] googletrans has MIT licence style, https://pypi.org/project/googletrans/

[6] SRL-BERT can be used non-exclusively https://allenai.org/terms

Matched predicates [took ---> opère]

[ARG1-A] --------> [ARG1-Un]
[ARG1-turning] --------> [ARG1-tournant]
[ARG1-point] --------> [ARG1-tournant]
[V-took] --------> [V-opère]
[ARG2-place] --------> [V-opère]
[ARGM-TMP-when] --------> [ARGM-TMP-lorsque]
[ARGM-TMP-Russia] --------> [ARGM-TMP-Russie]
[ARGM-TMP-concentrates] --------> [ARGM-TMP-concentre]
[ARGM-TMP-troops] --------> [ARGM-TMP-troupes]
[ARGM-TMP-on] --------> [ARGM-TMP-à]
[ARGM-TMP-the] --------> [ARGM-TMP-la]
[ARGM-TMP-Ukrainian] --------> [ARGM-TMP-ukrainienne]
[ARGM-TMP-border] --------> [ARGM-TMP-frontière]

Matched predicates [concentrates ---> concentre]

A --------> Un
turning --------> tournant
point --------> tournant
took --------> opère
place --------> opère
[ARGM-TMP-when] --------> [ARGM-TMP-lorsque]
[ARG0-Russia] --------> [ARG0-Russie]
[V-concentrates] --------> [V-concentre]
[ARG2-troops] --------> [ARG2-troupes]
[ARG1-on] --------> [ARG1-à]
[ARG1-the] --------> [ARG1-la]
[ARG1-Ukrainian] --------> [ARG1-ukrainienne]
[ARG1-border] --------> [ARG1-frontière]

Figure 2: Visualization of intermediate predicate-argument representation. Labels are specific to a given predicate.

distributed through a public github repository[7] and built on BERT, a transformer based model (Devlin et al., 2018). Word-to-word alignment maps source tokens to target tokens (see Fig. 3).

[ARG0-I] --------> [ARG0-Je]

[V-understand] --------> [V-comprends]

[ARG1-your] --------> [ARG1-votre]

[ARG1-dismay] --------> [ARG1-désarroi]

Figure 3: One-to-one mapping

In figure 3, Two sentences *I understand your dismay* and *Je comprends votre désarroi* are mapped one-to-one for each token. However, alignment does not always guarantee one-to-one mapping since the translation process produces divergences (Dorr, 1994) that make aligning languages difficult.

3.4 First Come, First Assign (FCFA) with Explainability

To address divergent SRL label projection, FCFA leverages isomorphic properties at the semantic level, i.e., similar structures for both source and target sentences, to handle divergence categories. Fig. 2 presents two intermediate representations for the two predicate-argument structures in the given sentence. Our approach transfers SRL labels from source (English) to target (French) based on these predicate-argument structures.

We also take advantage of SRL labels being specific to a given predicate. This ensures that SRL labels of aligned words are transferred successfully within divergence cases. For example, in Fig. 4, *is enough* is aligned with *suffit* which has a different POS. Although *is* and *enough* are aligned to *suffit* at the same time, FCFA assigns VERB to *suffit* since *[V-is]* is the first SRL label from the English side.

Matched predicates [is ---> suffit]

[ARG1-It] --------> [ARG1-Il]
[V-is] --------> [V-suffit]
[ARG2-enough] --------> [V-suffit]
[ARG2-to] --------> [ARG2-de]
[ARG2-pass] --------> [ARG2-passer]
[ARG2-specialized] --------> [ARG2-spécialisés]
[ARG2-software] --------> [ARG2-logiciels]
[ARG2-which] --------> [ARG2-qui]
[ARG2-all] --------> [ARG2-tous]
[ARG2-indicate] --------> [ARG2-indiquent]
[ARG2-less] --------> [ARG2-moins]
[ARG2-than] --------> [ARG2-de]
[ARG2-5] --------> [ARG2-5]
[ARG2-%] --------> [ARG2-%]

French Sentence: Il **suffit** d'ailleurs de passer des logiciels spécialisés qui indiquent tous moins de 5%
English Sentence: It **is enough** to pass specialized software which all indicate less than 5%

Figure 4: Categorial Divergence

Explainability of our approach is demonstrated through our visualization tool. We present a set of linguistic representations, one for each predicate followed by semantic role assignments. This representation provides two key pieces of information: (a) how we accommodate three divergence types for transferring SRL labels from the source dataset, and (b) interpretability of transferring SRL outputs

Fig. 4, 5, and 6 present projection from source to target, and show how our approach addresses cross-lingual divergences. Here, our representation tool reveals the presence of catego-

[7] https://github.com/andreabac3/Word_Alignment_BERT

Matched predicates [trust ---> confiance]
---
read --------> lit
the --------> les
programs --------> programmes
in --------> en
detail --------> détail
and --------> et
[ARG0-we] --------> [ARG0-on]
[V-trust] --------> [V-confiance]
" --------> "
[ARG1-believe] --------> [V-confiance]
" --------> "
French Sentence: lit les programmes en détail
et on **fait confiance** "Croire"
English Sentence: read the programs in detail
and we **trust** "believe"

Figure 5: Light Verb Divergence

Matched predicates [prevent ---> éviter]

But --------> Mais
its --------> son
program --------> programme
is --------> est
[ARG0-the] --------> [ARG0-le]
[ARG0-only] --------> [ARG0-seul]
[ARG0-one] --------> [ARG0-seul]
to --------> pour
[V-prevent] --------> [V-éviter]
[ARG1-France] --------> [ARG1-France]
[ARG2-from] --------> [ARG2-d']
[ARG2-going] --------> [ARG2-aller]
[ARG2-straight] --------> [ARG2-droit]
[ARG2-into] --------> [ARG2-dans]
[ARG2-the] --------> [ARG0-le]
[ARG2-wall] --------> [ARG2-mur]
French Sentence: Mais son programme est le seul
pour **éviter à la** France d'aller droit dans le mur
English Sentence: But its program is the only one to
prevent France from going straight into the wall

Figure 6: Structural Divergence

rial, light verb, and structural divergences (Dorr et al., 2002). Fig. 5 presents light verb divergence. Black colored words are not associated with *trust*/*confiance*, and only yellow/blue words are tied to the *trust*/*confiance* predicate-argument structure. Our approach assigns VERB to *confiance* and does not assign SRL label to *fait* since there is no word that is mapped to *fait*. Fig. 6 shows structural divergence, *prevent* is translated to *éviter à*, but only *éviter* has a SRL label. No SRL label is mapped to *à* as *à* is not aligned with an English word.

4 Experiments and Analyses

We describe our experiments for training and testing with projected french corpora and compare the performance of XSRL and FCFA.

4.1 Experiment setup

Two different corpora are projected from English using both XSRL[8] and FCFA. The result is SRL labeled French corpora (on Twitter and Wikipedia data) used to train French SRL models.

The datasets are converted into the format required for by AllenNLP's SRL-BERT training tool. Twitter data includes 25,549 sentences, and the Wikipedia dataset consists of 25,502 sentences. Both datasets are divided into train (80%), dev (18%), and test (2%) datasets. Thus, there are four models to train: XSRL-twitter, XSRL-wikipedia, FCFA-twitter, FCFA-wikipedia. We train our models using a learning rate of 5e-5 for the hugging-face_adamw optimizer and batch size of 32.

4.2 Experiment results and analyses

Our test data do not have ground truth SRL labels, so we define the ground truth to evaluate the model's performance. Therefore, we create an English corpus to use for ground truth. Analyses for accuracy uses precision (P), recall (R), and F1.

We define true positive (TP), false positive (FP), true negative (TN), and false negative (FN) as follows: A TP occurs when an English predicted SRL label and a French predicted SRL label are the same for aligned tokens. In Table 1, the predicate *was* and *était* are aligned and have the same SRL labels.

An FP occurs when an English token has no SRL label (O), and the corresponding French token has

[8]XSRL is Apache License 2.0 https://github.com/Heidelberg-NLP/xsrl_mbert_aligner/blob/main/LICENSE

a SRL label or when the English and French tokens have different SRL labels. In Table 1, *growing* and *grandie* are aligned, but *growing* has no SRL label and *grandie* is assigned a VERB. An example of a case with different SRL labels is the case of English *less* (ARG2) and French *moins* (ARG1).

A TN happens when both the English and French tokens are unassigned (O), as in *not* and *pas*. A FN occurs when an English token has a SRL label and the corresponding French token has no SRL label, e.g., *His* (ARG0) vs. *Son* (O) in Table 1.

Component	Example
True Positive	[V-was] -> [V-était]
False Positive	[O-growing] -> [V-grandie] [ARG2-less] -> [ARG1-moins]
True Negative	[O-not] -> [O-pas]
False Negative	[ARG0-His] ->[O-Son]

Table 1: Example of TP, FP, TN, FN

We explore the performance of two projection-based models: XSRL and FCFA. Table 2 and 3 present the results using 511 sentences (test dataset) from both Twitter and Wikipedia.

Table 2 presents XSRL projection performance. The models yield a higher F1 score (Twitter-Twitter 50.8, Wikipedia-Wikipedia 59.2) when the train and test data are from the same genre than when the train and test datasets are derived from different genres. F1 score drops when the training and test datasets come from different genres. Specifically, Wikipedia-trained models have a bigger drop in F1 score (-18.7%), but Twitter-trained models experience a lower drop in F1 score (-14.3%).

In Table 2, XSRL-Twitter-trained model shows a comparable decline in precision (-14.9%) and recall (-14%) when train and test dataset come from different genres. An XSRL-Wikipedia-trained model presents a higher decrement in recall (-25%), whereas decrement in precision is significantly lower (-0.06%).

Training Data	Test Data	P	R	F1
Twitter	Twitter	67.6	40.7	50.8
Twitter	Wikipedia	57.5	35.0	43.5
Wikipedia	Twitter	64.9	38.2	48.1
Wikipedia	Wikipedia	69.5	51.6	59.2

Table 2: Performance of XSRL projection models

The results of the FCFA projection-based model are shown in Table 3. The XSRL-projection-based model achieves higher F1 scores on same genre of train and test datasets. FCFA-projection-based model also yields higher F1 scores when the training and test data are generated from the same genres (Twitter-Twitter 50.8, Wikipedia-Wikipedia 59.6). When train data and test data differ, the FCFA-Wikipedia-trained model exhibits a more notable fall in F1 score (-20.4%) than that of a FCFA-Twitter-trained model (-14.9%).

In Table 3, precision and recall for both the Twitter and Wikipedia models decrease when the test dataset are not sourced from the same genre. The FCFA-Twitter-trained model presents a higher reduction for precision (-20.2%) than the reduction for recall (-11.9%) when test data sources are not from the same genre. The recall loss in the FCFA-Wikipedia-trained model is more notable (-29.8%) than the precision loss (-0.04%).

Training Data	Test Data	P	R	F1
Twitter	Twitter	69.4	40.1	50.8
Twitter	Wikipedia	55.5	35.3	43.2
Wikipedia	Twitter	64.8	37.3	47.4
Wikipedia	Wikipedia	67.8	53.2	59.6

Table 3: Performance of FCFA projection models

FCFA projection has significantly better time complexity: 107% faster than XSRL. We run at 2.2GHz on an Intel (R) Xeon (R) CPU on Google Colab, testing 511 sentences from Wikipedia and Twitter. The P, R, F1 for FCFA projection models appears comparable to XSRL on both Twitter and Wikipedia test data. The difference between each model is less than 1% in both Twitter and Wikipedia. Thus we have achieved explainable transferability of SRL labels with better speed performance and comparable P, R, F1 performance.

Model	Data	Time (secs)
FCFA	Twitter	**111.74**
(Daza and Frank, 2020)	Twitter	231.93
FCFA	Wikipedia	**204.71**
(Daza and Frank, 2020)	Wikipedia	424.91

Table 4: Speed Performance of FCFA and XSRL

5 Limitations

Although our approach provides an explainable projection method for cross-lingual semantic role labeling, there are three major limitations that need to be addressed: (a) our approach uses a transformer-

based model to align English and French; (b) evaluation data has low reliability since it is not a gold standard dataset; (c) training the French SRL system is less explainable. Our method relies on two key components: cross-lingual alignment based on mBERT and FCFA. Our approach enables explanations that elucidate how and why projection of SRL labels works, it does not explain how English words are mapped to French words since alignment uses a transformer-based model. We mitigate this problem by identifying word alignments using our visualization tool. These representations offer the validation of alignment outputs which is likely to help improve word alignments.

We expand cross-lingual SRL research to new genres using online social media data, and we create new corpora that contain SRL labels. However, these datasets do not have ground truth for SRL labels, making the model evaluation difficult. Even if we generate ground truth using SRL-BERT by AllenNLP, SRL-BERT trains on a different genre of data, and its performance still needs to be improved. More importantly, the SRL-BERT model is a transformer-based model which works through *black box* algorithms. Thus, this process is less interpretable, and we cannot explain how English words get SRL labels–we only explain how those roles are transferred to the non-English language.

Another limitation is inherent in the application of AllenNLP's SRL model trainer to create a French SRL model. This model uses BERT for training, which makes the inner workings of training on our new bilingual corpora opaque. Since our transferring of SRL labels are explainable, we can make refinements to FCFA, train and run SRL iteratively to see more immediate impacts of the refinements.

6 Conclusions and Future Work

We present an explainable projection-based cross-lingual SRL, which improves explainability of projection through representation tools. Our approach has three primary contributions, it (a) provides explainability as it shows each labels projection through visualization tools, which present decision making for which labels get projected and elucidates possible errors, (b) word-to-word alignment along with FCFA offers 107% faster than XSRL and equally accurate SRL label projection, and (c) a French SRL model yielding comparable performance to XSRL projection-based model

Future work will focus on improving explainability and the model's performance for the French SRL system. Our approach still operates partially using *black box* algorithms. For alignment, we need to enhance the explainability of how it aligns two languages, not just for French but also for other languages where hand-annotated labels are not readily available. It requires explicating how BERT operates in alignment and SRL tasks.

To improve the model's performance, using human-labeled data would be the best way to produce a gold-standard dataset. However, the human labeling process comes at the expense of time consumption. We can use human-labeled data only for development sets to achieve more accurate training outputs without significant loss of time complexity.

Our French SRL system-generated data can be used for training data as a silver dataset since our experiments yield meaningfully performance (F1 50.8) even if our dataset size is much smaller (708k words) than what SRL-BERT model has (2.9M words) and SRL-BERT performs 86.49 on its dataset. Furthermore, as an extension to this implementation, experiments will also be directed towards exploring possibilities of using FCFA annotation scheme in Abstract Meaning Representation (AMR) parsing in the multilingual domain.

Another future work will be correcting some of the incorrect transferred SRL labels. For example, in our system, when our input is *Mary attends school* in English and transfer its SRL labels to *Marie va à l'école*, our system transfers ARG1 to *à l'école* instead of ARG4 (see Fig. 7). The divergence here is tied to the variation between the verbs *go* and *attend*. Although (b) retains the same semantic meaning with (a), we need to correct SRL labels. This issue can be potentially identified by a tree based visualization as shown in Fig. 7.

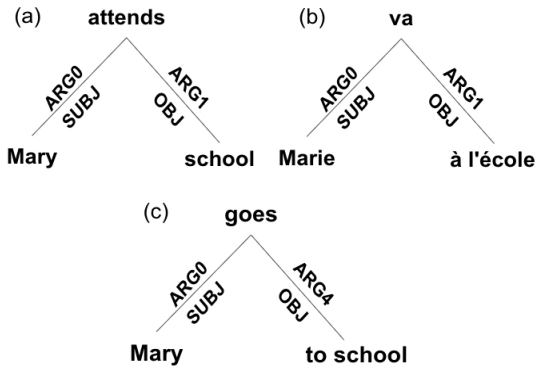

Figure 7: Dependency isomorphism

7 Ethical Considerations

Our French Twitter data are from Kaggle's open research Twitter dataset (Daignan, 2017) (https://www.kaggle.com/datasets/jeanmidev/french-presidential-election). User names can be found in this Twitter dataset, however, this Twitter dataset is open and publicly available. Our experiments also use Wikipedia data to expand our experimental evaluation; these are open to the public and freely usable.[9]

Potential risk of this work is bias in data. Our data are biased towards the social media genre. Care must be taken to generalize our performance. We mitigate the social media bias by combining it with a second data genre (i.e. Wikipedia). Another consideration to take into account is the risk that the alignment projection (Bacciu, 2021) is an unexplainable part of our approach, and also, alignment projection will not always produce accurate outputs. We reduce this risk by adding explainable approaches for alignment projection (Danilevsky et al., 2020) to our SRL model building pipeline. Our interpretable approach contributes to identifying inaccurate alignment and revealing the reason for the inaccurate outputs so that it can be easily remedied.

Acknowledgements

This research is based upon work supported by Defense Advanced Research Projects Agency (DARPA) under Contract No. HR001121C0186. Any opinions, findings and conclusions or recommendations expressed in this research are those of the authors and do not necessarily reflect the views of the Defense Advanced Research Projects Agency (DARPA)

References

Alan Akbik, Laura Chiticariu, Marina Danilevsky, Yunyao Li, Shivakumar Vaithyanathan, and Huaiyu Zhu. 2015. Generating High Quality Proposition Banks for Multilingual Semantic Role Labeling. *ACL-IJCNLP 2015 - 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing, Proceedings of the Conference*, 1:397–407.

[9]Wikipedia applies Creative Commons Attribution-ShareAlike 3.0 Unported License and GNU Free Documentation License

Andrea Bacciu. 2021. Cross-Lingual and Multilingual Word Alignment.

A Björkelund, L Hafdell, P Nugues Proceedings of the Thirteenth, and undefined 2009. 2009. Multilingual semantic role labeling. *aclanthology.org*, pages 43–48.

Sabine Buchholz and Erwin Marsi. 2006. CoNLL-X shared task on Multilingual Dependency Parsing. pages 149–164.

Jean-Michel Daignan. 2017. French presidential election: Extract from twitter about the french election,kaggle data set.

M Danilevsky, K Qian, R Aharonov, Y Katsis arXiv preprint arXiv . . . , and undefined 2020. 2020. A survey of the state of explainable AI for natural language processing. *arxiv.org*.

Marina Danilevsky, Kun Qian, Ranit Aharonov, Yannis Katsis, and Prithviraj Sen. 2021. Quantifying Explainability in NLP and Analyzing Algorithms for Performance-Explainability Tradeoff.

Angel Daza and Anette Frank. 2019. Translate and Label! An Encoder-Decoder Approach for Cross-lingual Semantic Role Labeling. *EMNLP-IJCNLP 2019 - 2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing, Proceedings of the Conference*, pages 603–615.

Angel Daza and Anette Frank. 2020. X-SRL: A parallel cross-lingual semantic role labeling dataset. *EMNLP 2020 - 2020 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference*, pages 3904–3914.

Jacob Devlin, Ming Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference*, 1:4171–4186.

Bonnie J Dorr. 1994. Machine Translation Divergences: A Formal Description and Proposed Solution.

Bonnie J. Dorr, Lisa Pearl, Rebecca Hwa, and Nizar Habash. 2002. DUSTer: A method for unraveling cross-language divergences for statistical word-level alignment. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2499:31–43.

Hao Fei, Meishan Zhang, and Donghong Ji. 2020a. Cross-Lingual Semantic Role Labeling with High-Quality Translated Training Corpus. pages 7014–7026.

Hao Fei, Meishan Zhang, Fei Li, and Donghong Ji. 2020b. Cross-lingual Semantic Role Labeling with Model Transfer. *IEEE/ACM Transactions on Audio Speech and Language Processing*, 28:2427–2437.

Boni García, Micael Gallego, Francisco Gortázar, and Mario Munoz-Organero. 2020. A Survey of the Selenium Ecosystem. *Electronics 2020, Vol. 9, Page 1067*, 9(7):1067.

Jonas Gehring, Michael Auli, David Grangier, Denis Yarats, and Yann N. Dauphin. 2017. Convolutional Sequence to Sequence Learning. *34th International Conference on Machine Learning, ICML 2017*, 3:2029–2042.

Pierre-Etienne Genest and Guy Lapalme. 2011. Framework for Abstractive Summarization using Text-to-Text Generation. pages 64–73.

SuHun Han. 2015. Googletrans: Free and Unlimited Google translate API for Python.

Hany Hassan, Anthony Aue, Chang Chen, Vishal Chowdhary, Jonathan Clark, Christian Federmann, Xuedong Huang, Marcin Junczys-Dowmunt, William Lewis, Mu Li, Shujie Liu, Tie-Yan Liu, Renqian Luo, Arul Menezes, Tao Qin, Frank Seide, Xu Tan, Fei Tian, Lijun Wu, Shuangzhi Wu, Yingce Xia, Dongdong Zhang, Zhirui Zhang, and Ming Zhou. 2018. Achieving Human Parity on Automatic Chinese to English News Translation.

Luheng He, Kenton Lee, Mike Lewis, and Luke Zettlemoyer. 2017. Deep Semantic Role Labeling: What Works and What's Next. *ACL 2017 - 55th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference (Long Papers)*, 1:473–483.

Shexia He, Zuchao Li, and Hai Zhao. 2019. Syntax-aware multilingual semantic role labeling. *EMNLP-IJCNLP 2019 - 2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing, Proceedings of the Conference*, pages 5350–5359.

R Hwa, P Resnik, A Weinberg, C Cabezas Natural language . . . , and undefined 2005. 2005. Bootstrapping parsers via syntactic projection across parallel texts. *cambridge.org*, 11(3):311–325.

M Kozhevnikov, I Titov Proceedings of the 51st Annual Meeting of, and undefined 2013. 2013. Cross-lingual transfer of semantic role labeling models. *aclanthology.org*, pages 1190–1200.

Katherine Lee, Orhan Firat, Ashish Agarwal, Clara Fannjiang, and David Sussillo. 2018. Hallucinations in Neural Machine Translation.

Z Li, S He, J Cai, Z Zhang, H Zhao, G Liu Proceedings of the . . . , and undefined 2018. 2018. A unified syntax-aware framework for semantic role labeling. *aclanthology.org*.

Ding Liu and Daniel Gildea. 2010. Semantic Role Features for Machine Translation.

Hui Liu, Qingyu Yin, and William Yang Wang. 2020. Towards explainable NLP: A generative explanation framework for text classification. *ACL 2019 - 57th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference*, pages 5570–5581.

Diego Marcheggiani, Anton Frolov, and Ivan Titov. 2017. A Simple and Accurate Syntax-Agnostic Neural Model for Dependency-based Semantic Role Labeling. *CoNLL 2017 - 21st Conference on Computational Natural Language Learning, Proceedings*, pages 411–420.

Brodie Mather, Bonnie J Dorr, Adam Dalton, William de Beaumont, Owen Rambow, and Sonja M. Schmer-Galunder. 2022. From Stance to Concern: Adaptation of Propositional Analysis to New Tasks and Domains.

Brodie Mather, Bonnie J. Dorr, Owen Rambow, and Tomek Strzalkowski. 2021. A General Framework for Domain-Specialization of Stance Detection. *The International FLAIRS Conference Proceedings*, 34.

Ryan McDonald, Joakim Nivre, Yvonne Quirmbach-Brundage, Yoav Goldberg, Dipanjan Das, Kuzman Ganchev, Keith Hall, Slav Petrov, Hao Zhang, Oscar Täckström, Claudia Bedini, Bertomeu Castelló, and Jungmee Lee. 2013. Universal Dependency Annotation for Multilingual Parsing.

Sanket Vaibhav Mehta, Jay Yoon Lee, and Jaime Carbonell. 2018. Towards semi-supervised learning for deep semantic role labeling. *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, EMNLP 2018*, pages 4958–4963.

A. More, Özlem Çetinoğlu, Çagri Çöltekin, Nizar Habash, B. Sagot, Djamé Seddah, Dima Taji, and Reut Tsarfaty. 2018. CoNLL-UL: Universal Morphological Lattices for Universal Dependency Parsing.

Phoebe Mulcaire, Swabha Swayamdipta, and Noah A. Smith. 2018. Polyglot semantic role labeling. *ACL 2018 - 56th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference (Long Papers)*, 2:667–672.

S Padó, M Lapata Journal of Artificial Intelligence Research, and undefined 2009. 2009. Cross-lingual annotation projection for semantic roles. *jair.org*, 36:307–340.

O Prazák, M Konopík RANLP, and undefined 2017. 2017. Cross-Lingual SRL Based upon Universal Dependencies. *acl-bg.org*.

Michael Roth and Mirella Lapata. 2016. Neural Semantic Role Labeling with Dependency Path Embeddings. *54th Annual Meeting of the Association for Computational Linguistics, ACL 2016 - Long Papers*, 2:1192–1202.

Y Shen, C Chu, F Cromieres Proceedings of the First . . . , and undefined 2016. 2016. Cross-language Projection of Dependency Trees with Constrained Partial Parsing for Tree-to-Tree Machine Translation. *aclanthology.org*, 1:1–11.

Peng Shi and Jimmy J. Lin. 2019. Simple BERT Models for Relation Extraction and Semantic Role Labeling. *ArXiv*.

Anders Søgaard. 2021. Book Review Explainable Natural Language Processing. 51:9781636392158.

David Yarowsky and Grace Ngai. 2001. Inducing Multilingual POS Taggers and NP Bracketers via Robust Projection Across Aligned Corpora.

M Zhang, H Jiang, A Aw, H Li, CL Tan Proceedings of ACL-08 . . . , and undefined 2008. 2008. A tree sequence alignment-based tree-to-tree translation model. *aclanthology.org*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2019. BERTScore: Evaluating Text Generation with BERT.