English-French Verb Phrase Alignment in Europarl for Tense Translation Modeling

Loaiciga Sanchez, Sharid; Meyer, Thomas; Popescu-Belis, Andréi

English-French Verb Phrase Alignment in Europarl for Tense Translation Modeling

Sharid Loáiciga*, Thomas Meyer†, Andrei Popescu-Belis†

*LATL-CUI, University of Geneva
Route de Drize 7
1227 Carouge, Switzerland
sharid.loaiciga@unige.ch

†Idiap Research Institute
Rue Marconi 19
1920 Martigny, Switzerland
{tmeyer, apbelis}@idiap.ch

Abstract

This paper presents a method for verb phrase (VP) alignment in an English/French parallel corpus and its use for improving statistical machine translation (SMT) of verb tenses. The method starts from automatic word alignment performed with GIZA++, and relies on a POS tagger and a parser, in combination with several heuristics, in order to identify non-contiguous components of VPs, and to label the aligned VPs with their tense and voice on each side. This procedure is applied to the Europarl corpus, leading to the creation of a smaller, high-precision parallel corpus with about 320 000 pairs of finite VPs, which is made publicly available. This resource is used to train a tense predictor for translation from English into French, based on a large number of surface features. Three MT systems are compared: (1) a baseline phrase-based SMT; (2) a tense-aware SMT system using the above predictions within a factored translation model; and (3) a system using oracle predictions from the aligned VPs. For several tenses, such as the French *imparfait*, the tense-aware SMT system improves significantly over the baseline and is closer to the oracle system.

Keywords: machine translation, verb tenses, verb phrase alignment

1. Introduction

The precise alignment of verb phrases (VPs) in parallel corpora is an important prerequisite for studying translation divergences in terms of tense-aspect-mode (TAM) as well as for modeling them computationally, in particular for Machine Translation (MT). In this paper, we present a method for aligning English and French verb phrases in the Europarl corpus, along with a quantitative study of tense mapping between these languages. The resulting resource comprises more than 300 000 pairs of aligned VPs with their tenses, and is made publicly available. Using the resource, we train a tense predictor for EN/FR translation and combine its output with the Moses phrase-based statistical MT system within a factored model. This improves the translation of VPs with respect to a baseline system. Moreover, for some tenses, our tense-aware MT system is closer to an oracle MT system (which has information of the correct target tense from our corpus) than to the baseline system.

The paper is organized as follows. We present related work on verb tenses in MT in Section 2. We introduce our high-precision VP alignment technique in Section 3 and analyzed the obtained resource quantitatively in Section 4, in terms of EN/FR tense mappings. We put our resource to use in Section 5 to train an automatic tense predictor, which we combine with a statistical MT system in Section 6, measuring the improvement of verb translation and of the overall BLEU score.

2. Related Work on Verb Tense Translation

Verb phrases (VPs) situate the event to which they refer in a particular time, and express its level of factuality along with the speaker's perception of it (Aarts, 2011). These tense-aspect-modality (TAM) characteristics are encoded quite differently across languages. For instance, when translating VPs into a morphologically rich language from a less rich one, mismatches of the TAM categories arise. The difficulties of generating highly inflected Romance VPs from English ones have been noted for languages such as Spanish (Vilar et al., 2006) and Brazilian Portuguese (Silva, 2010).

Research in statistical MT (SMT) only recently started to consider such verb tense divergences as a translation problem. For EN/ZH translation, given that tense is not morphologically marked in Chinese, Gong et al. (2012) built an n-gram-like sequence model that passes information from previously translated main verbs onto the next verb, with overall quality improvements o f up to 0.8 BLEU points. Ye et al. (2007) used a classifier to insert appropriate Chinese aspect markers which could also be used for EN/ZH translation.

Gojun and Fraser (2012) trained a phrase-based SMT system using POS-tags as disambiguation labels concatenated to English words which corresponded to the same German verb. This system gained up to 0.09 BLEU points over a system without the POS-tags.

For EN/FR translation, Grisot and Cartoni (2012) have shown that the English present perfect and simple past tenses may correspond to either *imparfait*, *passé composé* or *passé simple* in French and have identified a "narrativity" feature that helps to make the correct translation choice. Using an automatic classifier for narrativity, Meyer et al. (2013) showed that EN/FR translation of VPs in simple past tense was improved by 10% in terms of tense choice and 0.2 BLEU points. In this paper, we build on this idea and label English VPs directly with their predicted French tense for SMT.

English	French	VP EN	Tense EN	VP FR	Tense FR
I regret this since we are having to take action because others **have** not **done** their job.	Je le déplore car nous devons agir du fait que d'autres n'**ont** pas **fait** leur travail	have_done	present_perfect, active	ont_fait	passé_composé, active
To this end, I would like to remind you of the resolution of 15 September, which **recommended** that the proposal be presented as soon as possible.	En ce sens, je vous rappelle la résolution du 15 septembre, laquelle **recommandait** que la proposition soit présentée dans les plus brefs délais.	recommended	simple_past, active	recommandait	imparfait, active

Figure 1: Two sentences with one VP each (in bold) annotated with tense and voice on both English and French sides.

3. Method for VP Phrase Alignment

Our goal is to align verb phrases from the English and French sides of the Europarl corpus of European Parliament debates (Koehn, 2005), and to annotate each with VP labels indicating their tense, mode, and voice (active or passive) in both languages. The targeted annotation is exemplified in Figure 1 on two sentences with one VP each. The automatic procedure proposed here discards the pairs for which incoherent labels are found (as defined below), with the aim of selecting an unbiased, high-precision parallel corpus, which can be used for studies in corpus linguistics or for training automatic classifiers.

The following software is used to align and analyze VPs on both the English and French sides of Europarl:

- GIZA++ (Och and Ney, 2003) is used to retrieve word alignments between the two languages;
- a dependency parser (Henderson et al., 2008) is used for parsing the English side;
- Morfette (Chrupała et al., 2008) is used for French lemmatization and morphological analysis.

First, the parallel corpus is word-aligned using GIZA++ and each language is analyzed independently. From the parsing of the English sentences we retain the position, POS tags, heads and the dependency relation information. For the French side, we use both the morphological tags and the lemmas produced by Morfette. The three outputs are thereupon combined into a single file which contains the English parsing aligned to the French analysis according to the alignment produced by GIZA++.

In a second processing stage we use a set of hand-written rules to infer VPs and tense labels on the basis of the above annotations, independently for both sides of the parallel corpus. For example, if two words tagged as MD (Modal) and VB (Verb Base-form) are found, several tests follow: first, we check if MD is the head of VB, and then if they are bound by the VC (Verb Chain) dependency relation. If this is the case, then the sequence (MD VB) is interpreted as a valid VP. Last, in this particular case, the first word is tested to disambiguate between a future tense (the first word is *will* or *shall*) or a conditional (the first word is *should, would, ought, can, could, may*, or *might*).

The voice – active or passive – is determined for both languages, because it helps to distinguish between tenses with a similar syntactical configuration in French (e.g., *Paul est parti* vs. *Paul est menacé*, meaning 'Paul has left' vs. 'Paul is threatened'). Indeed, in French all forms of passive voice use the auxiliary ÊTRE (EN: *to be*), but a small set of intransitive verbs also use it in their compound past tense – these are essentially movement verbs and are recognized by our rules through a fixed list of lemmas. This example also illustrates the main reason for using Morfette for French parsing: it produces both morphological tagging and lemmatization, which are essential for determining the French tense.

We have defined 26 voice/tense combinations in English and 26 in French (13 active and 13 passive forms). Therefore, we have defined a set of 26 rules for each language, to recognize each tense and voice in the annotated VPs. Moreover, one rule was added in French for compound tenses with the auxiliary ÊTRE mentioned above.

At the end of the process, only pairs of aligned VPs assigned a valid tense both in English and French are retained.

4. Results of EN/FR VP Alignment

4.1. Quality Assessment

A set of 423 235 sentences from the Europarl English-French corpus (Koehn, 2005) was processed.[1] From this set, 3 816 sentences were discarded due to mismatches between the outputs of the parser and Morfette, leaving 419 419 annotated sentences. In total, 673 844 total English VPs were identified.

However, our focus is on verb tenses, therefore we discarded "non-finite" forms such as infinitives, gerunds and past particles acting as adjectives and kept only finite verbs (finite heads) – the full list of selected labels is given in the first column of Table 1. We selected 454 890 finite VPs (67.5%) and discarded 218 954 non-finite ones (32.5%).

Then, for each English VP with a tense label, we considered whether the French-side label was an acceptable one (erroneous labels are due to alignment mistakes and French lemmatization and morphological analysis mistakes). Table 1 shows the number of VPs for each English tense label, as well as the number of pairs with an acceptable label on the French side (number and percentage). On average about 81% of the pairs are selected at this stage. Overall, our method thus preserves slightly more than half of the input VP pairs ($67.5\% \times 81\%$), but ensures that both sides of the verb pair have acceptable labels.

To estimate the precision of the annotation (and noting that the above figure illustrates its "recall" rate), we evaluated manually a set of 413 VP pairs sampled from the final set, in terms of the accuracy of the VP boundaries and of the VP labels on each side. The results are presented in Table 2. The bottom line is that almost 90% of VP pairs have correct English and French labels, although not all of them

[1] A technical limitation of the parser prevented us from annotating the entire set of 2 008 710 sentences from the English-French section of Europarl, as intended.

have perfect VP boundaries. However, for corpus linguistics studies and even for use in MT, partially correct boundaries are not a major problem.

English tense	EN labels	FR labels	%
Simple past	52 198	39 475	76%
Past perfect	1 898	1 520	80%
Past continuous	1 135	878	77%
Past perfect continuous	31	26	84%
Present	270 145	219 489	81%
Present perfect	49 041	43 433	89%
Present continuous	22 364	19 118	86%
Present perfect continuous	1 104	979	89%
Future	17 743	12 963	73%
Future perfect	167	133	80%
Future continuous	675	546	81%
Future perfect continuous	1	1	100%
Conditional constructions	38 383	28 577	74%
Total	454 890	367 138	81%

Table 1: Number of annotated finite VPs for each tense category in the 419 419 sentences selected from Europarl.

	VP boundaries		Tense labels	
	EN	FR	EN	FR
Correct	97%	80%	95%	87%
Incorrect	1%	4%	5%	13%
Partial	2%	16%	–	–

Table 2: Human evaluation of the identification of VP boundaries and of tense labeling over 413 VP pairs.

4.2. Observations on EN/FR Tense Translation

We now examine the implications of our findings in terms of EN/FR verb tense translation. From Table 1, it appears that the proportion of VP pairs which had an acceptable French tense label is quite variable, reflecting the imperfections of precise alignment and the correctness of the analysis done by Morfette. The overwhelming disparity between the quantity of present tense (both in English and French) and all of the other tenses is to be noted: this tense alone represents about 60% of all finite VPs.

In fact, regarding French tense labeling, manual inspection revealed a rather systematic error with the identification of *conditional* and *future* tenses by Morfette: the pre-trained model we used appears to insert non-existent lemmas for these two tenses. We found that 1 490 out of 2 614 conditional verbs (57%) and 794 out of the 4 901 future tense verbs (16%) had similar errors which prevented them from receiving an acceptable tense label. Thus, in order to restrain any misleading input to the classifiers as well as any incorrect conclusion from the corpus study, we decided to remove the sentences containing any form of these two particular tenses, creating a subset of 203 140 sentences which was used in the subsequent translation experiments.

The final cleaned subset has a total of 322 086 finite VPs, which represent 70.8% of the total shown in Table 1. This means that almost 30% of correctly annotated sentences in English were discarded due to the mis-identification of French *future* or *conditional* modal.

Table 3 shows the distribution of tenses in the EN/FR parallel corpus, given as the number of occurrences and the percentage. These figures, which can be interpreted in both directions (EN/FR or FR/EN), show how a given source tense (or mode) can be translated into the target language, generally with several possibilities being observed for each tense. In fact, this distribution of tenses between English and French reveals a number of serious ambiguities of translation. The past tenses in particular – boldfaced in Table 3 – present important divergencies of translation, significant at $p < 0.05$. For example, the English present perfect (see the seventh column) can be translated into French either with a *passé composé* (61% of pairs), a *présent* (34%) or a *subjonctif* (2%). Similarly, the English simple past can be translated either by a *passé composé* (49% of pairs), or by a *présent* (25%), or by an *imparfait* (21%). This partially confirms the insights of the earlier study by Grisot and Cartoni (2012) using a corpus of 435 manually-annotated sentences.

5. Predicting EN/FR Tense Translation

One of the possible uses of the VP alignment described above is to train and to test an automatic tense predictor for EN/FR translation (keeping in mind when testing that the alignment is not 100% accurate). The hypothesis that we test is that, since such a predictor has access to a larger set of features than a SMT system, then when the two are combined, the translation of VPs and in particular of their tenses is improved. In this section, we present our tense predictor, and combine it with an MT system in the next section.

For predicting French tense automatically, we used the large gold-standard training set listed above (Section 4), using 196 140 sentences for training and 4 000 for tuning, and performing cross-validation. Therefore, when testing the combined system, the "test" set is made of fully unseen data.

We use a maximum entropy classifier from the Stanford Maximum Entropy package (Manning and Klein, 2003), with the features described hereafter (Subsection 5.1) and with different sets of French tenses as classes in order to maximize performance for the automatic translation task. In Subsection 5.2 we present results from experiments with various subsets of English features and various French tense classes in order to find the most valuable predictions for an MT system.

5.1. Features for Tense Prediction

We have used insights from previous work on classifying narrativity (Meyer et al., 2013) to design a similar feature set, but extended some of the features as we here have an up to 9-way[2] classification problem instead of just a binary one (narrative vs. non-narrative). We extract features from a series of parsers that were run on the English side of our data.

[2]All four *future* and *conditional* tenses from the original 13 tenses listed in Table 1 were grouped together into one single class. Details are given in Section 5.2.

French	English								
	Past continuous	Past perfect continuous	Past perfect	Present continuous	Present perfect continuous	Present perfect	Present	Simple past	Total
Imparfait	462 54%	7 27%	**365** **24%**	146 1%	18 2%	463 1%	1 510 1%	**8 060** **21%**	11 031 3%
Impératif				37 0%	1 0%	6 0%	203 0%	11 0%	258 0%
Passé composé	139 16%	2 8%	**214** **14%**	282 1%	325 33%	**26 521** **61%**	1253 1%	**19 402** **49%**	48 138 15%
Passé récent			1 0%	8 0%	3 0%	187 0%	2 0%	3 0%	204 0%
Passé simple	4 1%		6 0%	16 0%	2 0%	54 0%	42 0%	374 1%	498 0%
Plus-que-parfait	27 3%	8 31%	**782** **52%**	2 0%	4 0%	217 1%	22 0%	1 128 3%	2 190 1%
Présent	216 25%	9 35%	102 7%	18 077 96%	617 63%	**14 736** **34%**	211 334 97%	**9 779** **25%**	254 870 79%
Subjonctif	15 2%		28 2%	258 1%	6 1%	**1 053** **2%**	2 969 1%	568 1%	4 897 2%
Total	863 100%	26 100%	1 498 100%	18 826 100%	976 100%	43 237 100%	217 335 100%	39 325 100%	322 086 100%

Table 3: Distribution of the translation labels for 322 086 VPs in 203 140 annotated sentences. A blank cell indicates that no pairs were found for the respective combination, while a value of 0% indicates fewer than 1% of the occurrences. The values in bold indicate significant translation ambiguities.

We do not base our features on any parallel data and do not extract French features as we assume that we only have new and unseen English text at translation testing time. The three parsers are: (1) a dependency parser from Henderson et al. (2008); the Tarsqi toolkit for TimeML parsing (Verhagen and Pustejovsky, 2008); and (3) Senna, a syntactical parsing and semantic role labeling system based on convolutional neural networks (Collobert et al., 2011). From their output, we extract the following features:

Verb word form. The English verb to classify as it appears in the text.

Neighboring verb word forms. We not only extract the verb to classify, but also all other verbs in the current sentence, thus building a "bag-of-verbs". The value of this feature is a chain of verb word forms as they appear in the sentence.

Position. The numeric word index position of the verb in the sentence.

POS tags. We concatenate the POS tags of all occurring verbs, i.e. all POS tags such as VB, VBN, VBG, etc., as they are generated by the dependency parser. As an additional feature, we also concatenate all POS tags of the other words in the sentences.

Syntax. Similarly to POS tags, we get the syntactical categories and tree structures for the sentences from Senna.

English tense. Inferring from the POS tag of the English verb to classify, we apply a small set of rules as in Section 3 above to obtain a tense value out of the following possible attributes output by the dependency parser: VB (infinitive), VBG (gerund), VBD (verb in the past), and VBN (past participle).

Temporal markers. With a hand-made list of 66 temporal discourse markers we detect whether such markers are present in the sentence and use them as bag-of-word features.

Type of temporal markers. In addition to the actual marker word forms, we also consider whether a marker rather signals synchrony or asynchrony, or may signal both (e.g. *meanwhile*).

Temporal ordering. The TimeML annotation language tags events and their temporal order (FUTURE, INFINITIVE, PAST, PASTPART, etc.) as well as verbal aspect (PROGRESSIVE, PERFECTIVE, etc.). We thus use these tags obtained automatically from the output of the Tarsqi toolkit.

Dependency tags. Similarly to the syntax trees of the sentences with verbs to classify, we capture the entire dependency structure via the above-mentioned dependency parser.

Semantic roles. From the Senna output, we use the semantic role tag for the verb to classify, which is encoded in the standard IOBES format and can e.g. be of the form S-V or I-A1, indicating respectively head verb (V) of the sentence (S), or a verb belonging to the patient (A1) in between a chunk of words (I).

After analyzing the impact of the above features on a MaxEnt model for predicting French tenses, we noted poor performance when trying to automatically predict the *imparfait* (a past tense indicating a continuing action) and *sub-*

jonctif (a verb mode indicating some form of obligation). Because these two tenses are also among the most difficult to generate by a baseline SMT system, we added specific features to better predict these two tenses, aiming at improving also their translation from English. Both features were inferred from the analysis of examples extracted from a development set, hence already annotated for French verb tense.

Features for *imparfait*. We use a short hand-made list of potential indicators that an English simple past should be translated to *imparfait*: relative pronouns; relative clauses starting with *who, what, which, where*, or *why*; adverbs such as *repeatedly, constantly*; combinations of the verb *said* with the prepositions *that, as*.

Features for *subjonctif*. We first constructed a short list of French verbs that are likely to require in the next verb phrase the presence of the *subjonctif* mode if followed by a complement clause (clause starting by the FR conjunction *que*, roughly meaning *to* or *that* in English). Such verbs often express non-existing or hypothetical states or events, for instance, *souhaiter, espérer* or *supposer* (EN: *to wish, to hope* or *to suppose*). Besides these verbs, there are other expressions that can trigger the *subjonctif* in French in their subordinated clauses, such as the following ones: *so ... that*, or *delighted, clear, vision, way, good, expect, except, pleased, forward* followed by *to* or *that*. As our features can only come from the English source text, we built a list of English verbs and expressions that could likely require subjunctive mode in French.

5.2. Results on Tense Prediction

For our experiments, we used the set of 454 890 VP pairs described in Table 1. However, as noted in Section 4.2, the *future* and the *conditional* were often wrongly labeled by Morfette, so we decided to group these occurrences into a class labeled as '*other*', and keep only the 8 target French tenses shown in Table 3 plus the *passé antérieur* as output classes for the MaxEnt classifier. Keeping all sentences in the data but using the '*other*' class ensures that the classifier will have maximal coverage when unseen sentences are processed. This configuration of the classifier is referred to as 9_CLASSES+OTHER in what follows.
As the 'other' class is very frequent, we also experimented with a MaxEnt model that did not include this class, and was not trained on data containing it. This configuration of the classifier is referred to as 9_CLASSES in what follows.

MaxEnt Configuration	F1 (c.-v. data)	F1 (test set)
9_CLASSES+OTHER	0.75	n/a
9_CLASSES	0.85	0.83
9_CLASSES_EXT	0.85	0.83

Table 4: Performance of the MaxEnt models on predicting FR tenses. Reported are the micro-averaged F1 scores for different model configurations and data sets.

After having evaluated these models, we decided that 9_CLASSES was the most suitable in order to reach the highest number of correctly predicted tense labels to be used for SMT. In addition, as explained in the section above, we extended the MaxEnt model with two specific features to better predict the *imparfait* tense and the *subjonctif* mode. We thus extended the features of the 9_CLASSES system (with all French tenses except the '*other*' class) into a final classifier called 9_CLASSES_EXT.
The classification results for the three different systems are shown in Table 4. F1 scores are given for 10-fold cross-validation on the entire training set and, when relevant (i.e. when considered for the translation task), also on the test set.

	9_CLASSES		9_CLASSES_EXT	
French tense	F1 (cv)	F1 (test)	F1 (cv)	F1 (test)
Imparfait	0.48	0.40	0.47	0.44
Passé composé	0.77	0.73	0.76	0.72
Impératif	0.29	0.00	0.24	0.00
Passé simple	0.16	0.00	0.09	0.00
Plus-que-parfait	0.55	0.36	0.51	0.25
Présent	0.92	0.91	0.91	0.91
Subjonctif	0.33	0.16	0.29	0.17
Passé récent	0.16	0.00	0.22	0.00
Macro-average	0.46	0.32	0.44	0.31

Table 5: Performance of two MaxEnt models on predicting specific French tenses in terms of F1 scores per class for 10-fold cross-validation and on the test set. The macro-average is giving equal weight to each class, not considering their distribution. The French tenses not occurring in the test set are not listed.

The scores show that the large '*other*' class has a detrimental influence on the overall classification performance, likely because it distorts precision on the small classes (such as *passé antérieur* or *impératif*). When this class is removed, performance reaches up to 0.85 F1 score, which is the highest observed value.
However, we also performed an analysis per tense class, showing F1 scores for each class in Table 5 for the 9_CLASSES and the 9_CLASSES_EXT models. The second model, using also the features for better predicting *imparfait* and *subjonctif*, does not appear to improve in cross-validation performance; still, on the test set, the two tenses have slight gains of respectively 0.04 and 0.01 F1 scores. In the following section, we test both classifier configurations for their effect on tense-aware SMT systems.

6. Tense-aware Statistical MT System

Aiming at assessing the usefulness of the annotated corpus in a MT context, we trained three systems: a baseline system used for control comparison; a tense-aware system built using the automatically predicted tenses; and third, a tense-aware oracle system which serves as an indicator of the maximal improvement we can expect if all translations of tenses were correctly predicted. Henceforth we will refer to these systems as "baseline", "predicted" and "oracle" respectively.
We used the Moses toolkit (Koehn et al., 2007) for the three systems with phrase-based translation models; in addition, for the predicted and oracle systems, we used factored translation models (Koehn and Hoang, 2007), which

allow for integration of arbitrary linguistic markup (i.e., factors) at the word level. There can be several factors per word such as lemmas, POS tags, etc. In our case however, as we wanted to check translation improvement due to verb tense only, the verbs in the sentences receive one tense label from the tense predictor or as it is given in the oracle annotated datasets (e.g. 'was|IMP' for *imparfait*), and all other words are set to the '|null' factor. The three systems were built by partitioning the total 203 140 sentences as follows: 196 140 sentences for training; 4 000 sentences for tuning; and 3 000 sentences for testing.

We evaluated the 3 000 test sentences of the corpus using different automatic metrics and using manual error inspection as well. The BLEU and METEOR scores (Papineni et al., 2002; Denkowski and Lavie, 2011) obtained are given in Table 6. It can be noted that the oracle system gained 0.5 points over the baseline for the BLEU score, while the predicted system gained 0.12 points. This amount was rather stable after each of the three completed tunings.

The METEOR score shows a positive difference of 0.0029 points between the baseline and the oracle system and a minimal negative difference of 0.0005 points between the baseline and the predicted system. Since this score is calculated not only on the basis of exact matches but also on stems, the small difference means that only few verb stems are changed. This is the expected behavior since a tense-aware system should mainly modify inflectional suffixes, but not the stems. The negative difference of the predicted system could indicate a tendency to change the lexical choice of a verb's translation, even when the tense may be correct (cf. manual evaluation scores for lexical choice in Table 9).

System	BLEU				METEOR
	Run 1	Run 2	Run 3	Average	
Baseline	27.73	27.63	27.64	27.67	0.4912
Predicted	27.83	27.75	27.78	27.79	0.4907
Oracle	28.23	28.15	28.13	28.17	0.4941

Table 6: BLEU and METEOR scores after tuning.

	Baseline	Oracle	Predicted	#sent.
Imparfait	24.10	25.32	24.57	122
Passé composé	29.80	30.82	30.08	359
Impératif	19.08	19.72	18.70	4
Passé simple	**13.34**	**16.15**	**14.09**	6
Plus-que-parfait	**21.27**	**23.44**	**23.22**	17
Présent	27.55	27.97	27.59	2618
Subjonctif	26.81	27.72	26.07	78
Passé recent	**24.54**	**30.50**	**30.08**	3

Table 7: BLEU scores per expected French tense for the three systems. Largest score increases are boldfaced. The number of sentences for each class is given in the last column.

The increment of BLEU is still quite significant, as the detailed BLEU scores presented in Table 7 reveal. Indeed, when each expected French tense is observed in detail, it is evident that the model works particularly well with the less frequent tenses. In other words, high-frequency tenses such as the present tense, which do not have virtually any translation ambiguity from English to French – as evidenced by the 97.24% of this tense translated as French *Présent* tense in Table 3 – tend to hide (in the overall scores) the genuine improvement of the tense-aware systems on ambiguous tenses.

Table 7 also shows that the oracle system obtained improved results throughout all the tenses, with the *passé simple, plus-que-parfait* and *passé récent* doing much better than the baseline. The predicted model improves over the baseline as well, for most French tenses, especially for *plus-que-parfait* and *passé récent*, for which it nearly reaches the oracle performance level. Only for *subjonctif* and *impératif* the performance falls below the baseline system, due to poor classifier performance for these two tenses.

A qualitative assessment of the systems was done by means of a detailed manual evaluation of 313 sentences, comprising 654 VPs, from the test set. The results are shown in Table 8. Each annotated VP was evaluated in three different categories. "TAM" refers to the Tense, Aspect and Mode features – our main evaluation interest. "Lexical choice" assesses the correctness of the verbal lemma, this criterion captures the cases in which the TAM features of a VP were improved but the lemma itself changed, being then penalized by BLEU. Finally, "Agreement" refers to whether a translation is free from errors of person and number agreement. For the first two categories, we evaluated if the translation was different than the reference yet correct ($\neq$ ref) or identical (= ref).

In terms of tense translation the oracle model outperformed the baseline by an average of 24% and up to 27%, while the predicted system outperformed the baseline by an average of 10%. The ratio of these results is within our expectations: the predicted system is in between the upper bound of the oracle system and the lower bound of the baseline system. Concerning the Lexical choice and the Agreement categories, they did not change much between the three systems. When looking at the results per French translated tense (Table 9) we confirmed that low-frequency verbs are better translated by both tense-aware systems, for instance the *passé simple* and the *passé récent*.

On the other hand, the *imparfait* and the *subjonctif* tenses (boldfaced in Table 9) reveal that English tenses with a real translation ambiguity were better translated by the tense aware systems. For instance, while most of the *present perfect* English VPs were translated as *passé composé* by the baseline – since this is the most frequent translation with up to 61% of the instances according to the translation distribution given in Table 3, the tense aware models boosted the instantiation of the *imparfait* tense in French.

Concerning the predicted model, for the *imparfait* tense in particular, it can be noted that the results are closer to the oracle than to the baseline as evidenced by the boldfaced counts in Table 9; however, when it comes to the *subjonctif* tense, its results are closer to the baseline. This observation demonstrates that the predictor results have a direct impact on the MT results and confirms that our method has a meaningful effect on the translation of verb tenses.

In Figure 2 we present an example taken from the test set. The first verb is incorrectly translated with a French *infini-*

System	TAM			Lexical choice			Agreement		Total VPs
	Incorrect	Correct		Incorrect	Correct		Incorrect	Correct	
		$\neq$ ref.	= ref.		$\neq$ ref.	= ref.			
Baseline	206	61	387	47	267	340	118	536	654
	32%	9%	59%	7%	41%	51%	18%	82%	100%
Predicted	146	79	429	50	255	349	140	514	654
	22%	12%	66%	8%	39%	53%	21%	79%	100%
Oracle	52	39	563	60	247	347	122	532	654
	8%	6%	86%	9%	38%	53%	19%	81%	100%

Table 8: General results of the manual evaluation of 313 sentences from the test set.

French tense	System	TAM			Lexical choice			Agreement		Total VPs
		Incorrect	Correct		Incorrect	Correct		Incorrect	Correct	
			$\neq$ ref	= ref		$\neq$ ref	= ref			
Imparfait	Baseline	**82**	**15**	**41**	7	56	75	27	111	138
	Predicted	**42**	**23**	**73**	14	55	69	29	109	
	Oracle	**13**	**4**	**121**	14	51	73	27	111	
Passé composé	Baseline	28	6	129	14	68	81	32	131	163
	Predicted	31	10	122	10	68	85	53	110	
	Oracle	14	5	144	8	66	89	32	131	
Présent	Baseline	21	20	201	16	93	133	34	208	242
	Predicted	12	18	212	14	81	147	26	216	
	Oracle	12	19	211	13	87	142	25	217	
Subjonctif	Baseline	**63**	**11**	**6**	10	35	35	16	64	80
	Predicted	**51**	**17**	**12**	11	37	32	20	60	
	Oracle	**11**	**7**	**62**	20	29	31	24	56	

Table 9: Results of the manual evaluation given per expected tense. Only the most frequent tenses are presented.

tif by the baseline system, but correctly by the one using automatic tense predictions and the one using oracle tense labels. The second verb is also incorrectly translated into a *présent*, indicative mode, while a *subjonctif* was required. Although this is correctly generated by the oracle system, the predicted one has actually skipped the word. Of course, some of the surrounding words are also of variable correctness.

SOURCE	... that we support a system that is clearer than the current one ...
BASELINE	... que nous *soutenir* un système qui *est* plus claire que le système actuel ...
PREDICTED	... que nous soutenons un système ∅ plus claires que le système actuel ...
ORACLE	... que nous soutenons un système qui soit clair que ce que le programme actuel ...
REFERENCE	... que nous soutenons un système qui soit plus clair que le système actuel ...

Figure 2: Translations produced by the baseline vs. predicted vs. oracle systems along with source and reference.

7. Conclusion

We have proposed a fully automatic method for high precision VP alignment. Even though the method selects only about half of the verb phrases, the large number of occurrences that is available still ensures a large resource. Manual evaluation of a sample showed that about 90% of the labeled occurrences receive a correct label. Incorrect labels were due to the fact that the errors produced by each tool sum up: word-alignments (NULL or non-verbal), English-side parsing (mistakes in long-distance dependencies in compound forms), and French-side tagging (frequent mistakes on conditionals and even lemmas, for unclear reasons).

Based on the annotated corpus, we implemented a French tense predictor that is able to automatically learn and predict which French tense an English verb should be translated into. The results of this predictor were used in a factored SMT model whose results were compared to a baseline and an oracle system. We found that overall, our method improves the quality of verbal translations, increasing the general BLEU score up to 0.5 points.

8. Acknowledgments

This work was performed while the first author was at the Idiap Research Institute. We are grateful for the funding of this work to the Swiss National Science Foundation (SNSF), under its Sinergia program, grants n. CRSI22_127510 (COMTIS project, see `www.idiap.ch/project/comtis`) and n. CRSII2_147653 (MODERN project, see `www.idiap.ch/project/modern`). The resources described in this article are available at `https://www.idiap.ch/dataset/tense-annotation`.

9. References

Aarts, Bas. (2011). *Oxford Modern English Grammar*. Oxford University Press.

Chrupała, Grzegorz, Dinu, Georgiana, and van Genabith, Josef. (2008). Learning morphology with morfette. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC)*, Marrakech, Morocco.

Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., and Kuksa, P. (2011). Natural language processing (almost) from scratch. *Journal of Machine Learning Research (JMLR)*, 11:2461–2505.

Denkowski, Michael and Lavie, Alon. (2011). Meteor 1.3: Automatic metric for reliable optimization and evaluation of machine translation systems. In *Proceedings of the EMNLP 2011 Workshop on Statistical Machine Translation (WMT)*.

Gojun, Anita and Fraser, Alexander. (2012). Determining the placement of German verbs in English-to-German SMT. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, pages 726–735, Avignon, France.

Gong, Zhengxian, Zhang, Min, Tan, Chew Lim, and Zhou, Guodong. (2012). N-gram-based tense models for statistical machine translation. In *Proceedings of the Joint Conference on Empirical Methods in Natural Language Processing (EMNLP) and Computational Natural Language Learning (CoNLL)*, pages 276–285, Jeju Island, Korea.

Grisot, Cristina and Cartoni, Bruno. (2012). Une description bilingue des temps verbaux: étude contrastive en corpus. *Nouveaux cahiers de linguistique française*, 30:101–117.

Henderson, James, Merlo, Paola, Musillo, Gabriele, and Titov, Ivan. (2008). A latent variable model of synchronous parsing for syntactic and semantic dependencies. In *Proceedings of the 12th Conference on Computational Natural Language Learning (CoNLL)*, pages 178–182, Manchester, UK.

Koehn, Philipp and Hoang, Hieu. (2007). Factored translation models. In *Proceedings of the Joint Conference on Empirical Methods in Natural Language Processing (EMNLP) and Computational Natural Language Learning (CoNLL)*, pages 868–876, Prague, Czech Republic.

Koehn, Philipp, Hoang, Hieu, Birch, Alexandra, Callison-Burch, Chris, Federico, Marcello, Bertoli, Nicola, Cowan, Brooke, Shen, Wade, Moran, Christine, Zens, Richard, Dyer, Christopher J., Bojar, Ondřej, Constantin, Alexandra, and Herbst, Evan. (2007). Moses: Open Source Toolkit forStatistical Machine Translation. In *Proceedings of the 45th Annual Meeting of the ACL, Demo and Poster Sessions*, pages 177–180, Prague, Czech Republic.

Koehn, Philipp. (2005). Europarl: A parallel corpus for statistical machine translation. In *Proceedings of MT Summit X*.

Manning, Christopher and Klein, Dan. (2003). Optimization, MaxEnt models, and conditional estimation without magic. In *Tutorial at HLT-NAACL and 41st ACL conferences*, Edmonton, Canada and Sapporo, Japan.

Meyer, Thomas, Grisot, Cristina, and Popescu-Belis, Andrei. (2013). Detecting narrativity to improve English to French translation of simple past verbs. In *Proceedings of the 1st DiscoMT Workshop at ACL 2013 (51st Annual Meeting of the Association for Computational Linguistics)*, pages 33–42, Sofia, Bulgaria.

Och, Franz Josef and Ney, Hermann. (2003). A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29:19–51.

Papineni, Kishore, Roukos, Salim, Ward, Todd, and Zhu, Wei-Jing. (2002). BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics (ACL)*, pages 311–318.

Silva, Lucia. (2010). Fine-tuning in Brazilian Portuguese-English statistical transfer machine translation: Verbal tenses. In *Proceedings of the NAACL HLT 2010 Student Research Workshop*, pages 58–63, Los Angeles, CA.

Verhagen, Marc and Pustejovsky, James. (2008). Temporal processing with the TARSQI toolkit. In *Proceedings of the 22nd International Conference on Computational Linguistics (COLING), Demonstrations*, pages 189–192, Manchester, UK.

Vilar, David, Xu, Jia, D'Haro, Luis Fernando, and Ney, Hermann. (2006). Error analysis of statistical machine translation output. In *Proceedings of the Fith International Conference on Language Resources and Evaluation (LREC)*, Genoa, Italy.

Ye, Yang, Schneider, Karl-Michael, and Abney, Steven. (2007). Aspect marker generation for English-to-Chinese machine translation. In *Proceedings of MT Summit XI*, pages 521–527, Copenhagen, Denmark.