Detecting Narrativity to Improve English to French Translation of Simple Past Verbs

Thomas Meyer
Idiap Research Institute and EPFL
Martigny and Lausanne, Switzerland
thomas.meyer@idiap.ch

Cristina Grisot
University of Geneva
Switzerland
cristina.grisot@unige.ch

Andrei Popescu-Belis
Idiap Research Institute
Martigny, Switzerland
andrei.popescu-belis@idiap.ch

Abstract

The correct translation of verb tenses ensures that the temporal ordering of events in the source text is maintained in the target text. This paper assesses the utility of automatically labeling English Simple Past verbs with a binary discursive feature, narrative vs. non-narrative, for statistical machine translation (SMT) into French. The narrativity feature, which helps deciding which of the French past tenses is a correct translation of the English Simple Past, can be assigned with about 70% accuracy (F1). The narrativity feature improves SMT by about 0.2 BLEU points when a factored SMT system is trained and tested on automatically labeled English-French data. More importantly, manual evaluation shows that verb tense translation and verb choice are improved by respectively 9.7% and 3.4% (absolute), leading to an overall improvement of verb translation of 17% (relative).

1 Introduction

The correct rendering of verbal tenses is an important aspect of translation. Translating to a wrong verbal tense in the target language does not convey the same meaning as the source text, for instance by distorting the temporal order of the events described in a text. Current statistical machine translation (SMT) systems may have difficulties in choosing the correct verb tense translations, in some language pairs, because these depend on a wider-range context than SMT systems consider. Indeed, decoding for SMT is still at the phrase or sentence level only, thus missing information from previously translated sentences (which is also detrimental to lexical cohesion and co-reference).

In this paper, we explore the merits of a discourse feature called *narrativity* in helping SMT systems to improve their translation choices for English verbs in the Simple Past tense (henceforth, SP) into one of the three possible French past tenses. The narrativity feature characterizes each occurrence of an SP verb, either as *narrative* (for ordered events that happened in the past) or *non-narrative* (for past states of affairs). Narrativity is potentially relevant to EN/FR translation because three French past tenses can potentially translate an English Simple Past (SP), namely the *Passé Composé* (PC), *Passé Simple* (PS) or *Imparfait* (IMP). All of them can be correct translations of an EN SP verb, depending on its narrative or non-narrative role.

The narrativity feature can be of use to SMT only if it can be assigned with sufficient precision over a source text by entirely automatic methods. Moreover, a narrativity-aware SMT model is likely to make a difference with respect to baseline SMT only if it is based on additional features that are not captured by, e.g., a phrase-based SMT model. In this study, we use a small amount of manually labeled instances to train a narrativity classifier for English texts. The (imperfect) output of this classifier over the English side of a large parallel corpus will then be used to train a narrativity-aware SMT system. In testing mode, the narrativity classifier provides input to the SMT system, resulting (as we will show below) in improved tense and lexical choices for verbs, and a modest but statistically significant increase in BLEU and TER scores. Overall, the method is similar in substance to our previous work on the

Proceedings of the Workshop on Discourse in Machine Translation (DiscoMT), pages 33–42,
Sofia, Bulgaria, August 9, 2013.

combination of a classifier for discourse connectives with an SMT system (Meyer and Popescu-Belis, 2012; Meyer et al., 2012).

The paper is organized as follows. Section 2 exemplifies the hypothesized relation between narrativity and the translations of the English Simple Past into French, along with related work on modeling tense for MT. The automatic labeling experiments are presented in Section 3. Experiments with SMT systems are presented in Section 4, with results from both automatic (4.3) and manual translation scoring (4.4), followed by a discussion of results and suggestions on improving them (Section 5).

2 English Simple Past in Translation

2.1 Role of Narrativity: an Example

The text in Figure 1 is an example taken from the 'newstest 2010' data described in Section 4 below. In this four-sentence discourse, the English verbs, all in Simple Past, express a series of events having occurred in the past, which no longer affect the present. As shown in the French translation by a baseline SMT system (not aware of narrativity), the English SP verbs are translated into the most frequent tense in French, as learned from the parallel data the SMT was trained on.

When looking more closely, however, it appears that the Simple Past actually conveys different temporal and aspectual information. The verbs *offered* and *found* describe actual events that were ordered in time and took place subsequently, whereas *were* and *was* describe states of general nature, not indicating any temporal ordering.

The difference between narrative and non-narrative uses of the English Simple Past is not always captured correctly by the baseline SMT output in this example. The verbs in the first and third sentences are correctly translated into the French PC (one of the two tenses for past narratives in French along with the PS). The verb in the second sentence is also correctly rendered as IMP, in a non-narrative use. However, the verb *was* in the fourth sentence should also have been translated as an IMP, but from lack of sufficient information, it was incorrectly translated as a PC. A non-narrative label could have helped to find the correct verb tense, if it would have been annotated prior to translation.

EN: (1) After a party, I offered [**Narrative**] to throw out a few glass and plastic bottles. (2) But, on Kounicova Ulice, there were [**Non-narrative**] no colored bins to be seen. (3) Luckily, on the way to the tram, I found [**Narrative**] the right place. (4) But it was [**Non-narrative**] overflowing with garbage.

FR from BASELINE MT system: (1) Après un parti, j'**ai proposé** pour rejeter un peu de verre et les bouteilles en plastique. (2) Mais, sur Kounicova Ulice, il n'y **avait** pas de colored bins à voir. (3) Heureusement, sur la manière de le tramway, j'**ai trouvé** la bonne place. (4) Mais il ***a été** débordés avec des ramasseurs.

Figure 1: Example English text from 'newstest 2010' data with narrativity labels and a translation into French from a baseline SMT. The tenses generated in French are, respectively: (1) PC, (2) IMP, (3) PC, (4) PC. The mistake on the fourth one is explained in the text.

2.2 Modeling Past Tenses

The classical view on verb tenses that express past tense in French (PC, PS and IMP) is that both the PC and PS are perfective, indicating that the event they refer to is completed and finished (Martin, 1971). Such events are thus single points in time without internal structure. However, on the one hand, the PC signals an accomplished event (from the aspectual point of view) and thus conveys as its meaning the possible consequence of the event. The PS on the other hand is considered as aspectually unaccomplished and is used in contexts where time progresses and events are temporally ordered, such as narratives.

The IMP is imperfective (as its name suggests), i.e. it indicates that the event is in its preparatory phrase and is thus incomplete. In terms of aspect, the IMP is unaccomplished and provides background information, for instance ongoing state of affairs, or situations that are repeated in time, with an internal structure.

Conversely, in English, the SP is described as having as its main meaning the reference to past tense, and as specific meanings the reference to present or future tenses identified under certain contextual conditions (Quirk et al., 1986). Corblin and de Swart (2004) argue that the SP is aspectually 'transparent', meaning that it applies to

all types of events and it preserves their aspectual class.

The difficulty for the MT systems is thus to choose correctly among the three above-mentioned tenses in French, which are all valid possibilities of translating the English SP. When MT systems fail to generate the correct tense in French, several levels of incorrectness may occur, exemplified in Figure 2 with sentences taken from the data used in this paper (see Section 3 and Grisot and Cartoni (2012)).

1. In certain contexts, tenses may be quite interchangeable, which is the unproblematic case for machine translation, depending also on the evaluation measure. In Example 1 from Figure 2, the verb *étaient considérées* (were seen) in IMP has a focus on temporal length which is preserved even if the translated tense is a PC (*ont été considérées*, i.e. have been seen) thanks to the adverb *toujours* (always)).

2. In other contexts, the tense proposed by the MT system can sound strange but remains acceptable. For instance, in Example 2, there is a focus on temporal length with the IMP translation (*voyait*, viewed) but this meaning is not preserved if a PC is used (*a vu*, has viewed) though it can be recovered by the reader.

3. The tense output by an MT system may be grammatically wrong. In Example 3, the PC *a renouvelé* (has renewed) cannot replace the IMP *renouvelaient* (renewed) because of the conflict with the imperfective meaning conveyed by the adverbial *sans cesse* (again and again).

4. Finally, a wrong tense in the MT output can be misleading, if it does not convey the meaning of the source text but remains unnoticed by the reader. In Example 4, using the PC *a été* leads to the interpretation that the person was no longer involved when he died, whereas using IMP *était* implies that he was still involved, which may trigger very different expectations in the mind of the reader (e.g. on the possible cause of the death, or its importance to the peace process).

1. **EN**: Although the US viewed Musharraf as an agent of change, he has never achieved domestic political legitimacy, and his policies **were seen** as rife with contradictions.
FR: Si les Etats-Unis voient Moucharraf comme un agent de changement, ce dernier n'est jamais parvenu à avoir une légitimité dans son propre pays, où ses politiques ont toujours **été considérées** (PC) **/ étaient considérées** (IMP) comme un tissu de contradictions.

2. **EN**: Indeed, she even persuaded other important political leaders to participate in the planned January 8 election, which she **viewed** as an opportunity to challenge religious extremist forces in the public square.
FR: Benazir Bhutto a même convaincu d'autres dirigeants de participer aux élections prévues le 8 janvier, qu'elle **voyait** (IMP) **/ ?a vu** (PC) comme une occasion de s'opposer aux extrémistes religieux sur la place publique.

3. **EN**: The agony of grief which overpowered them at first, **was** voluntarily **renewed**, was sought for, was created again and again...
FR: Elles s'encouragèrent l'une l'autre dans leur affliction, la **renouvelaient** (IMP) **/** l'***a renouvelé** (PC) volontairement, et sans cesse...

4. **EN**: Last week a person who **was** at the heart of the peace process passed away.
FR: La semaine passée une personne qui **était** (IMP) **/ a été** (PC) au cœur du processus de paix est décédée.

Figure 2: Examples of translations of the English SP by an MT system, differing from the reference translation: (1) unproblematic, (2) strange but acceptable, (3) grammatically wrong (*), and (4) misleading.

2.3 Verb Tenses in SMT

Modeling verb tenses for SMT has only recently been addressed. For Chinese/English translation, Gong et al. (2012) built an n-gram-like sequence model that passes information from previously translated main verbs onto the next verb so that its tense can be more correctly rendered. Tense is morphologically not marked in Chinese, unlike in English, where the verbs forms are modified according to tense (among other factors). With such a model, the authors improved translation by up to 0.8 BLEU points.

Conversely, in view of English/Chinese translation but without implementing an actual translation system, Ye et al. (2007) used a classifier to generate and insert appropriate Chinese aspect markers that in certain contexts have to follow the Chinese verbs but are not present in the English source texts.

For translation from English to German, Gojun and Fraser (2012) reordered verbs in the English source to positions where they normally occur in

German, which usually amounts to a long-distance movement towards the end of clauses. Reordering was implemented as rules on syntax trees and improved the translation by up to 0.61 BLEU points.

In this paper, as SMT training needs a large amount of data, we use an automatic classifier to tag instances of English SP verbs with narrativity labels. The labels output by this classifier are then modeled when training the SMT system.

3 Automatic Labeling of Narrativity

3.1 Data

A training set of 458 and a test set of 118 English SP verbs that were manually annotated with narrativity labels (narrative or non-narrative) was provided by Grisot and Cartoni (2012) (see their article for more details about the data). The training set consists of 230 narrative and 228 non-narrative instances, the test set has 75 narrative instances and 43 non-narrative ones. The sentences come from parallel EN/FR corpora of four different genres: literature, news, parliamentary debates and legislation. For each instance, the English sentence with the SP verb that must be classified, as well as the previous and following sentences, had been given to two human annotators, who assigned a narrative or non-narrative label. To avoid interference with the translation into French, which could have provided clues about the label, the translations were not shown to annotators[1].

Annotators agreed over only 71% of the instances, corresponding to a $kappa$ value of only 0.44. As this is at the lower end of the acceptable spectrum for discourse annotation (Carletta, 1996), one of the important questions we ask in this paper is: what can be achieved with this quality of human annotation, in terms of an automatic narrativity classifier (intrinsic performance) and of its use for improving verb translation by SMT (extrinsic evaluation)? It must be noted that instances on which the two annotators had disagreed were resolved (to either narrative or non-narrative) by looking at the French human translation (an acceptable method given that our purpose here is translation into French), thus increasing the quality of the annotation.

[1]The goal was to focus on the narrativity property, regardless of its translation. However, annotations were adjudicated also by looking at the FR translation. For a different approach, considering exclusively the tense in translation, see the discussion in Section 5.

Model	Recall	Prec.	F1	κ
MaxEnt	0.71	0.72	**0.71**	$+$**0.43**
CRF	0.30	0.44	0.36	-0.44

Table 1: Performance of MaxEnt and CRF classifiers on narrativity. We report recall, precision, their mean (F1), and the $kappa$ value for class agreement.

3.2 Features for Narrativity

The manually annotated instances were used for training and testing a Maximum Entropy classifier using the Stanford Classifier package (Manning and Klein, 2003). We extracted the following features from the sentence containing the verb to classify and the preceding sentence as well, thus modeling a wider context than the one modeled by phrase-based SMT systems. For each verb form, we considered its POS tag and syntactical category, including parents up to the first verbal phrase (VP) parent node, as generated by Charniak and Johnson's constituent parser (2005). This parser also assigns special tags to auxiliary (AUX) and modal verbs (MD), which we include in the features.

We further used a TimeML parser, the Tarsqi Toolkit (Verhagen et al., 2005; Verhagen and Pustejovsky, 2008), which automatically outputs an XML-like structure of the sentence, with a hypothesis on the temporal ordering of the events mentioned. From this structure we extract event markers such as PAST-OCCURRENCE and aspectual information such as STATE.

Temporal ordering is often also signaled by other markers such as adverbials (e.g., *three weeks before*). We manually gathered a list of 66 such temporal markers and assigned them, as an additional feature, a label indicating whether they signal synchrony (e.g., *meanwhile, at the same time*) or asynchrony (e.g., *before, after*).

3.3 Results of Narrativity Labeling

With the above features, we obtained the classification performance indicated in Table 1. The MaxEnt classifier reached 0.71 F1 score, which is similar to the human annotator's agreement level. Moreover, the $kappa$ value for inter-class agreement was 0.43 between the classifier and the human annotation, a value which is also close to the $kappa$ value for the two human annotators. In a sense, the classifier thus reaches the highest scores

that are still meaningful, i.e. those of inter-coder agreement. As a baseline for comparison, the majority class in the test set (the 'narrative' label) would account for 63.56% of correctly classified instances, whereas the classifier correctly labeled 72.88% of all test instances.

For further comparison we built a CRF model (Lafferty et al., 2001) in order to label narrativity in sequence of other tags, such as POS. The CRF uses as features the two preceding POS tags to label the next POS tag in a sequence of words. The same training set of 458 sentences as used above was POS-tagged using the Stanford POS tagger (Toutanova et al., 2003), with the `left3words-distsim` model. We replaced the instances of 'VBD' (the POS tag for SP verbs) with the narrativity labels from the manual annotation. The same procedure was then applied to the 118 sentences of the test set on which CRF was evaluated.

Overall, the CRF model only labeled narrativity correctly at an F1 score of 0.36, while $kappa$ had a negative value signaling a weak inverse correlation. Therefore, it appears that the temporal and semantic features used for the MaxEnt classifier are useful and account for the much higher performance of MaxEnt, which is used in the SMT experiments described below.

We further evaluate the MaxEnt classifier by providing in Table 2 the confusion matrix of the automatically obtained narrativity labels over the test set. Labeling non-narrative uses is slightly more prone to errors (32.6% error rate) than narrative ones (24% errors), likely due to the larger number of narratives vs. non-narratives in the training and the test data.

	System		
Reference	**Narr.**	**Non-narr.**	**Total**
Narrative	57	18	75
Non-narr.	14	29	43
Total	71	47	118

Table 2: Confusion matrix for the labels output by the MaxEnt classifier (System) versus the gold standard labels (Reference).

4 SMT with Narrativity Labels

4.1 Method

Two methods to use labels conveying to SMT information about narrativity were explored (though more exist). First, as in our initial studies applied to discourse connectives, the narrativity labels were simply concatenated with the SP verb form in EN (Meyer and Popescu-Belis, 2012) – see Example 2 in Figure 3. Second, we used factored translation models (Koehn and Hoang, 2007), which allow for any linguistic annotation to be considered as additional weighted feature vectors, as in our later studies with connectives (Meyer et al., 2012). These factors are log-linearly combined with the basic features of phrase-based SMT models (phrase translation, lexical and language model probabilities).

To assess the performance gain of narrativity-augmented systems, we built three different SMT systems, with the following names and configurations:

- BASELINE: plain text, no verbal labels.
- TAGGED: plain text, all SP verb forms concatenated with a narrativity label.
- FACTORED: all SP verbs have narrativity labels as source-side translation factors (all other words labeled 'null').

1. BASELINE SMT: on wednesday the čssd declared the approval of next year's budget to be a success. the people's party was also satisfied.

2. TAGGED SMT: on wednesday the čssd declared-**Narrative** the approval of next year's budget to be a success. the people's party was-**Non-narrative** also satisfied.

3. FACTORED SMT: on wednesday the čssd declared|**Narrative** the approval of next year's budget to be a success. the people's party was|**Non-narrative** also satisfied.

Figure 3: Example input sentence from 'newstest 2010' data for three translation models: (1) plain text; (2) concatenated narrativity labels; (3) narrativity as translation factors (the '|null' factors on other words were omitted for readability).

Figure 3 shows an example input sentence for these configurations. For the FACTORED SMT model, both the EN source word and the factor

information are used to generate the FR surface target word forms. The tagged or factored annotations are respectively used for the training, tuning and test data as well.

For labeling the SMT data, no manual annotation is used. In a first step, the actual EN SP verbs to be labeled are identified using the Stanford POS tagger, which assigns a 'VBD' tag to each SP verb. These tags are replaced, after feature extraction and execution of the MaxEnt classifier, by the narrativity labels output by the latter. Of course, the POS tagger and (especially) our narrativity classifier may generate erroneous labels which in the end lead to translation errors. The challenge is thus to test the improvement of SMT with respect to the baseline, in spite of the noisy training and test data.

4.2 Data

In all experiments, we made use of parallel English/French training, tuning and testing data from the translation task of the Workshop on Machine Translation (`www.statmt.org/wmt12/`).

- For *training*, we used Europarl v6 (Koehn, 2005), original EN[2] to translated FR (321,577 sentences), with 66,143 instances of SP verbs labeled automatically: 30,452 are narrative and 35,691 are non-narrative.
- For *tuning*, we used the 'newstest 2011' tuning set (3,003 sentences), with 1,401 automatically labeled SP verbs, of which 807 are narrative and 594 non-narrative.
- For *testing*, we used the 'newstest 2010' data (2,489 sentences), with 1,156 automatically labeled SP verbs (621 narrative and 535 non-narrative).

We built a 5-gram language model with SRILM (Stolcke et al., 2011) over the entire FR part of Europarl. Tuning was performed by Minimum Error Rate Training (MERT) (Och, 2003). All translation models were phrase-based using either plain text (possibly with concatenated labels) or factored training as implemented in the Moses SMT toolkit (Koehn et al., 2007).

[2]We only considered texts that were originally authored in English, not translated into it from French or a third-party language, to ensure only proper tenses uses are observed. The relevance of this constraint is discussed for connectives by Cartoni et al. (2011).

4.3 Results: Automatic Evaluation

In order to obtain reliable automatic evaluation scores, we executed three runs of MERT tuning for each type of translation model. With MERT being a randomized, non-deterministic optimization process, each run leads to different feature weights and, as a consequence, to different BLEU scores when translating unseen data.

Table 3 shows the average BLEU and TER scores on the 'newstest 2010' data for the three systems. The scores are averages over the three tuning runs, with resampling of the test set, both provided in the evaluation tool by Clark et al. (2011) (`www.github.com/jhclark/multeval`). BLEU is computed using jBLEU V0.1.1 (an exact reimplementation of NIST's 'mteval-v13.pl' script without tokenization). The Translation Error Rate (TER) is computed with version 0.8.0 of the software (Snover et al., 2006). A t-test was used to compute p values that indicate the significance of differences in scores.

Translation model	BLEU	TER
BASELINE	21.4	61.9
TAGGED	21.3	61.8
FACTORED	21.6*	61.7*

Table 3: Average values of BLEU (the higher the better) and TER (the lower the better) over three tuning runs for each model on 'newstest 2010'. The starred values are significantly better ($p < 0.05$) than the baseline.

In terms of overall BLEU and TER scores, the FACTORED model improves performance over the BASELINE by +0.2 BLEU and -0.2 TER (as lower is better), and these differences are statistically significant at the 95% level. On the contrary, the concatenated-label model (noted TAGGED) slightly decreases the global translation performance compared to the BASELINE. A similar behavior was observed when using labeled connectives in combination with SMT (Meyer et al., 2012).

The lower scores of the TAGGED model may be due to the scarcity of data (by a factor of 0.5) when verb word-forms are altered by concatenating them with the narrativity labels. The small improvement by the FACTORED model of overall scores (such as BLEU) is also related to the scarcity of SP verbs: although their translation is

improved, as we will now show, the translation of all other words is not changed by our method, so only a small fraction of the words in the test data are changed.

4.4 Results: Human Evaluation

To assess the improvement specifically due to the narrativity labels, we manually evaluated the FR translations by the FACTORED model for the 207 first SP verbs in the test set against the translations from the BASELINE model. As the TAGGED model did not result in good scores, we did not further consider it for evaluation. Manual scoring was performed along the following criteria for each occurrence of an SP verb, by bilingual judges looking both at the source sentence and its reference translation.

- Is the narrativity label correct? ('correct' or 'incorrect') – this is a direct evaluation of the narrativity classifier from Section 3
- Is the verb tense of the FACTORED model more accurate than the BASELINE one? (noted '+' if improved, '=' if similar, '−' if degraded)
- Is the lexical choice of the FACTORED model more accurate than the BASELINE one, regardless of the tense? (again noted '+' or '=' or '−')
- Is the BASELINE translation of the verb phrase globally correct? ('correct' or 'incorrect')
- Is the FACTORED translation of the verb phrase globally correct? ('correct' or 'incorrect')

Tables 4 and 5 summarize the counts and percentages of improvements and/or degradations of translation quality with the systems FACTORED and BASELINE. The correctness of the labels, as evaluated by the human judges on SMT test data, is similar to the values given in Section 3 when evaluated against the test sentences of the narrativity classifier. As shown in Table 4, the narrativity information clearly helps the FACTORED system to generate more accurate French verb tenses in almost 10% of the cases, and also helps to find more accurate vocabulary for verbs in 3.4% of the cases. Overall, as shown in Table 5, the FACTORED model yields more correct translations of the verb phrases than the BASELINE in 9% of the cases – a small but non-negligible improvement.

Criterion	Rating	N.	%	Δ
Labeling	correct	147	71.0	
	incorrect	60	29.0	
Verb tense	+	35	17.0	+9.7
	=	157	75.8	
	−	15	7.2	
Lexical choice	+	19	9.2	+3.4
	=	176	85.0	
	−	12	5.8	

Table 4: Human evaluation of verb translations into French, comparing the FACTORED model against the BASELINE. The Δ values show the clear improvement of the narrativity-aware factored translation model.

System	Rating	Number	%
BASELINE	correct	94	45.5
	incorrect	113	54.5
FACTORED	correct	113	54.5
	incorrect	94	45.5

Table 5: Human evaluation of the global correctness of 207 translations of EN SP verbs into French. The FACTORED model yields 9% more correct translations than the BASELINE one.

An example from the test data shown in Figure 4 illustrates the improved verb translation. The BASELINE system translates the SP verb *looked* incorrectly into the verb *considérer* (*consider*), in wrong number and its past participle only (*considérés*, plural). The FACTORED model generates the correct tense and number (IMP, *semblait*, singular) and the better verb *sembler* (*look, appear*). This example is scored as follows: the labeling is correct ('yes'), the tense was improved ('+'), the lexical choice was improved too ('+'), the BASELINE was incorrect while the FACTORED model was correct.

5 Discussion and Future Work

When looking in detail through the translations that were degraded by the FACTORED model, some were due to the POS tagging used to find the EN SP verbs to label. For verb phrases made of an auxiliary verb in SP and a past participle (e.g. *was born*), the POS tagger outputs *was*/VBD *born*/VBN. As a consequence, our classifier only considers *was*, as non-narrative, although *was*

EN: tawa hallae **looked|Non-narrative** like many other carnivorous dinosaurs.

FR BASELINE: tawa hallae ***considérés** comme de nombreuses autres carnivores dinosaures.

FR FACTORED: tawa hallae **semblait** comme de nombreux autres carnivores dinosaures.

Figure 4: Example comparison of a baseline and improved factored translation. The '|null' factors in EN were omitted for readability. See the text for a discussion.

born as a whole is a narrative event. This can then result in wrong FR tense translations. For instance, the fragment *nelson mandela was*|Non-narrative *born on* ... is translated as: *nelson mandela *était né en* ..., which in FR is pluperfect tense instead of the correct Passé Composé *est né* as in the reference translation. A method to concatenate such verb phrases to avoid such errors is under work.

A further reason for the small improvements in translation quality might be that factored translation models still operate on rather local context, even when the narrativity information is present. To widen the context captured by the translation model, labeling entire verbal phrase nodes in hierarchical or tree-based syntactical models will be considered in the future. Moreover, it has been shown that it is difficult to choose the optimal parameters for a factored translation model (Tamchyna and Bojar, 2013).

In an alternative approach currently under work, a more direct way to label verb tense is implemented, where a classifier can make use of the same features as those extracted here (in Section 3.2), but its classes are those that directly indicate which target verb tense should be output by the SMT. Thus, not only SP verbs can be considered and no intermediate category such as narrativity (that is more difficult to learn) is needed. The classifier will predict which FR tense should be used depending on the context of the EN verbs, for which the FR tense label can be annotated as above, within a factored translation model. Through word alignment and POS tagging, this method has the additional advantage of providing much more training data, extracted from word alignment of the verb phrases, and can be applied to all tenses, not only SP. Moreover, the approach is likely to learn which verbs are preferably translated with which tense: for instance, the verb *started* is much more likely to become *a commencé* (PC) in FR than to *commençait* (IMP), due to its meaning of a punctual event in time, rather than a continuous or repetitive one.

6 Conclusion

The paper presented a method to automatically label English verbs in Simple Past tense with a binary pragmatic feature, narrativity, which helps to distinguish temporally ordered events that happened in the past ('narrative') from past states of affairs ('non-narrative'). A small amount of manually annotated data, combined with the extraction of temporal semantic features, allowed us to train a classifier that reached 70% correctly classified instances. The classifier was used to automatically label the English SP verbs in a large parallel training corpus for SMT systems. When implementing the labels in a factored SMT model, translation into French of the English SP verbs was improved by about 10%, accompanied by a statistically significant gain of +0.2 BLEU points for the overall quality score. In the future, we will improve the processing of verb phrases, and study a classifier with labels that are directly based on the target language tenses.

Acknowledgments

We are grateful for the funding of this work to the Swiss National Science Foundation (SNSF) under the COMTIS Sinergia Project, n. CRSI22_127510 (see `www.idiap.ch/comtis/`). We would also like to thank the anonymous reviewers for their helpful suggestions.

References

Jean Carletta. 1996. Assessing Agreement on Classification Tasks: The Kappa Statistic. *Computational Linguistics*, 22:249–254.

Bruno Cartoni, Sandrine Zufferey, Thomas Meyer, and Andrei Popescu-Belis. 2011. How Comparable are Parallel Corpora? Measuring the Distribution of General Vocabulary and Connectives. In *Proceedings of 4th Workshop on Building and Using Comparable Corpora (BUCC)*, pages 78–86, Portland, OR.

Eugene Charniak and Mark Johnson. 2005. Coarse-to-fine n-best Parsing and MaxEnt Discriminative

Reranking. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 173–180, Ann Arbor, MI.

Jonathan Clark, Chris Dyer, Alon Lavie, and Noah Smith. 2011. Better Hypothesis Testing for Statistical Machine Translation: Controlling for Optimizer Instability. In *Proceedings of ACL-HLT 2011 (46th Annual Meeting of the ACL: Human Language Technologies)*, Portland, OR.

Francis Corblin and Henriëtte de Swart. 2004. *Handbook of French Semantics*. CSLI Publications, Stanford, CA.

Anita Gojun and Alexander Fraser. 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.

Zhengxian Gong, Min Zhang, Chew Lim Tan, and Guodong Zhou. 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.

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

Philipp Koehn and Hieu Hoang. 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.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, and Evan Herbs. 2007. Moses: Open Source Toolkit for Statistical Machine Translation. In *Proceedings of 45th Annual Meeting of the Association for Computational Linguistics (ACL), Demonstration Session*, pages 177–180, Prague, Czech Republic.

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

John Lafferty, Andrew McCallum, and Fernando Pereira. 2001. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. *The Journal of Machine Learning Research*, 8:693–723.

Christopher Manning and Dan Klein. 2003. Optimization, MaxEnt Models, and Conditional Estimation without Magic. In *Tutorial at HLT-NAACL and 41st ACL conferences*, Edmonton, Canada and Sapporo, Japan.

Robert Martin. 1971. *Temps et aspect: essai sur l'emploi des temps narratifs en moyen français*. Klincksieck, Paris, France.

Thomas Meyer and Andrei Popescu-Belis. 2012. Using Sense-labeled Discourse Connectives for Statistical Machine Translation. In *Proceedings of the EACL 2012 Joint Workshop on Exploiting Synergies between IR and MT, and Hybrid Approaches to MT (ESIRMT-HyTra)*, pages 129–138, Avignon, FR.

Thomas Meyer, Andrei Popescu-Belis, Najeh Hajlaoui, and Andrea Gesmundo. 2012. Machine Translation of Labeled Discourse Connectives. In *Proceedings of the Tenth Biennial Conference of the Association for Machine Translation in the Americas (AMTA)*, San Diego, CA.

Franz Josef Och. 2003. Minimum Error Rate Training in Statistical Machine Translation. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 160–167, Sapporo, Japan.

Randolph Quirk, Sidney Greenbaum, Geoffrey Leech, and Jan Svartvik. 1986. *A Comprehensive Grammar of the English Language*. Pearson Longman, Harlow, UK.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A Study of Translation Edit Rate with Targeted Human Annotation. In *Proceedings of the Tenth Biennial Conference of the Association for Machine Translation in the Americas (AMTA)*, Cambridge, MA.

Andreas Stolcke, Jing Zheng, Wen Wang, and Victor Abrash. 2011. SRILM at Sixteen: Update and Outlook. In *Proceedings of the IEEE Automatic Speech Recognition and Understanding Workshop*, Waikoloa, Hawaii.

Aleš Tamchyna and Ondřej Bojar. 2013. No Free Lunch in Factored Phrase-Based Machine Translation. In *Proceedings of the 14th International Conference on Computational Linguistics and Intelligent Text Processing (CICLING)*, Samos, Greece.

Kristina Toutanova, Dan Klein, Christopher Manning, and Yoram Singer. 2003. Feature-rich Part-of-Speech Tagging with a Cyclic Dependency Network. In *Proceedings of Human Language Technology Conference and the North American Chapter of the Association for Computational Linguistics (HLT-NAACL)*, pages 252–259, Edmonton, CA.

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

Marc Verhagen, Inderjeet Mani, Roser Sauri, Jessica Littman, Robert Knippen, Seok Bae Jang, Anna Rumshisky, John Phillips, and James Pustejovsky. 2005. Automating Temporal Annotation with TARSQI. In *Proceedings of the 43th Annual Meeting of the Association for Computational Linguistics (ACL), Demo Session*, pages 81–84, Ann Arbor, USA.

Yang Ye, Karl-Michael Schneider, and Steven Abney. 2007. Aspect Marker Generation for English-to-Chinese Machine Translation. In *Proceedings of MT Summit XI*, pages 521–527, Copenhagen, Danmark.