Linguistic **I**ssues in **L**anguage **T**echnology – **LiLT**

Predicting and Using a Pragmatic Component of Lexical Aspect

of Simple Past Verbal Tenses for Improving English-to-French Machine Translation

Sharid Loáiciga
& Cristina Grisot

Published by CSLI Publications

LiLT volume 13, issue 3 August 2016

Predicting and Using a Pragmatic Component of Lexical Aspect

of Simple Past Verbal Tenses for Improving English-to-French Machine Translation

Sharid Loáiciga, *University of Geneva* & Cristina Grisot, *University of Geneva & University of Neuchâtel*

Abstract

This paper proposes a method for improving the results of a statistical Machine Translation system using *boundedness*, a pragmatic component of the verbal phrase's lexical aspect. First, the paper presents manual and automatic annotation experiments for lexical aspect in English-French parallel corpora. It will be shown that this aspectual property is identified and classified with ease both by humans and by automatic systems. Second, Statistical Machine Translation experiments using the *boundedness* annotations are presented. These experiments show that the information regarding lexical aspect is useful to improve the output of a Machine Translation system in terms of better choices of verbal tenses in the target language, as well as better lexical choices. Ultimately, this work aims at providing a method for the automatic annotation of data with *boundedness* information and at contributing to Machine Translation by taking into account linguistic data.

LiLT Volume 13, Issue 3, August 2016.
Predicting and Using a Pragmatic Component of Lexical Aspect.

1 Introduction

This paper aims at improving the results of Statistical Machine Translation (SMT) systems with respect to *verbal tenses*. Verbal tenses are the primary linguistic source of temporal reference, i.e. the localization of events and states in time. Verbal tenses express temporal location of eventualities[1] with respect to the moment of speech and with respect to each other. For example, the past perfect form in sentence (1) locates the two eventualities before S ('moment of speech'), hence in the past, and the eventuality of *hard working* before the moment when another eventuality, *the son's disappearance*, occurs.

(1) He *had worked* hard the day when his son disappeared.

The choice of the verbal tense depends on fine-grained temporal interpretations of the utterance in context. In this paper, we show the translation of verbal tenses can be improved at the sentence-level when we make use of a pragmatic component of the verbal phrase's lexical aspect, i.e. boundedness, and the way in which it is perceived by human speakers.

We have investigated the usefulness of boundedness for correctly translating the English Simple Past (SP) into French using a SMT system. The four most frequently used translations of the SP in French are: *passé composé* (PC), *imparfait* (IMP), *passé simple* (PS) and *présent* (PRES). Loáiciga et al. (2014) showed that this mapping has a skewed distribution in favour of the PC translation. Since SMT systems favour the most frequent translation and make little use of context or meaning, the other three possible translations are often not generated. This yields translations that can be in some cases ungrammatical and in other cases grammatical but not native-like. As an illustration, in Figure 1 we show a sentence translated into French by a baseline system built for our experiments (Section 5) together with its translation by Google Translate [2]. In the example, the verb in bold is an English SP verb translated using the French PC by both systems. However, the reference translation proposes a PRES form.

Our hypothesis is that boundedness is relevant to disambiguate French translations of the English SP. With this assumption, the main goal of this paper is to improve the MT of English SP verbs by integrating this temporal property into a SMT system.

[1] The term *eventuality* as well as *situation* are generic terms which include all aspectual types of verbs. The term *event* is often used as a synonym, especially in computational approaches. In the linguistic literature, however, the term *event* does not include verbs considered as *states*.

[2] https://translate.google.com/#en/fr/

Source	It **was** not uncommon for cattle-rustling to occur between cattle-keeping tribes.
Phrase-based SMT	Il **a été** une pratique courante pour vol de bétail lorsque l'on a affaire entre cattle-keeping tribus.
Google Translate	Il n'**a** pas **été** rare pour vol de bétail de se produire entre l'élevage du bétail tribus.
Reference	Les vols de bétail ne **sont** pas rares entre tribus d'éleveurs.

FIGURE 1 Example outputs of SMT systems.

Our method can be summarized in four main steps:

1. We annotate the English part of a parallel corpus with lexical aspect labels i.e., *bounded* or *unbounded*.
2. We train a classifier on the human annotated corpus for predicting lexical aspect labels.
3. We use the classifier to automatically annotate a large corpus with the lexical aspect labels.
4. We build a SMT system using the large automatically annotated corpus.

This paper is organized as follows. Section 2 introduces the general background of this research. Concretely, Section 2.1 discusses the categories aspect and tense and the operationalization of boundedness for the human annotation experiments. Sections 2.2 to 2.4 are dedicated to current research on discourse and MT, the automatic prediction of verbal tenses and MT of verbal tenses for different language pairs. Section 3 describes the corpus data and human annotation experiments. Section 4 presents the lexical aspect prediction tasks. In 4.1, we present a model for predicting the aspect labels, i.e., *bounded* or *unbounded* for each verb instance in the corpus; in 4.2, we present a second model for aspect prediction based exclusively on automatic features. Section 5 describes the annotation of a large corpus (using the classifier built in the previous section) used for training an aspect-aware SMT system. Specifically, two systems are built and compared: one which uses the aspect labels and one which does not. Finally, Section 6 concludes the paper.

2 Background

2.1 Aspect and Tense

In this section, we will discuss the role played by the categories of *tense*, *grammatical aspect* and *lexical aspect* for expressing temporal reference

in a natural language[3]. The category of tense is defined as *the grammatical marking of the localization of a situation in time* (Comrie, 1985) and the meaning of a verbal tensed form was first formalized by Reichenbach (1947). He proposed using three temporal coordinates (event point E, moment of speech S and reference point R) and two temporal relations (precedence and simultaneity). In Reichenbach's model, a verbal form such as the Past Perfect receives the following formalization $E < R < S$. Hence the meaning of the form 'past perfect' indicates that the moment when the event took place is previous to the reference moment which is previous to speech moment. This is temporal information provided by the category of tense.

Grammatical aspect is defined as *the grammatical marking of the speaker's viewpoint on the situation referred to* with respect to its internal consistency. It can be *perfective* or *imperfective* (Comrie, 1976). Specifically, the perfective aspect indicates that the situation should be viewed as a single whole, while the imperfective indicates that the speaker focused on the internal structure of the situation or on its progression. It is morphologically marked in Slavic languages, whose verbal systems are organised around this category. In English, only a subtype of the imperfective aspect is morphologically marked through the progressive *-ing* morpheme. As for Romance languages, they do not mark grammatical aspect morphologically. The progressive aspect may be expressed in some Romance languages, such as French, through the lexical form *être en train de*, but not in Romanian. In the literature, it is assumed by default that the IMP verbal tense expresses the imperfective aspect whereas the PS and the PC verbal tenses express the perfective aspect. However, studies in pragmatics challenged this assumption suggesting that they only have imperfective and perfective contextual usages.

Lexical aspect is defined as *the semantic category that refers to the temporal information inherent to the VP* (containing the verb, which is the head of the phrase, and its internal arguments) (Dowty, 1979, Comrie, 1976, Depraetere, 1995b). Temporal information provided by lexical aspect is independent of the speaker's way of viewing the situation and by category of tense. The most well known and used aspectual classification comes from Vendler (1957). Vendler suggested a four branch distinction: *states* ('love', 'know'), *activities* ('run', 'push a

[3]Natural languages are classified as *tensed* and *tenseless* languages. Tensed languages are further classified as *tense prominent*, such as Germanic and Romance languages, and as *aspect prominent*, such as Slavic languages, whereas tenseless languages (such as Mandarin Chinese) use other means to express temporal reference, for instance lexical aspect, temporal and aspectual particles and adverbials.

cart)', *accomplishments* ('run a mile', 'draw a circle') and *achievements* ('recognise', 'reach the top'). Lexical aspect is also called *ontological* aspect because it refers to ontological features used to describe situations, such as stativity, dynamicity, homogeneity, durativity, agentivity and telicity among others[4]. Vendler's classification was suggested for the English verbs and makes use of the homogeneity and telicity ontological features (states and activities are homogeneous and atelic whereas accomplishments and achievements are non-homogeneous and telic).

However, in many cases, tense and grammatical aspect modify and override inherent temporal features of a situation. For example, the example in (2) shows that an activity such as *run the marathon* that normally is homogeneous and atelic becomes a non-homogeneous and telic eventuality (i.e. an accomplishment) because it is expressed through the habitual aspect (that is a sub-part of the imperfective aspect, see Comrie (1976)).

(2) For years, I used to run the marathon in two hours and a half but now it takes three.

An issue regarding the interpretation of eventualities related to telicity is *boundedness* (Declerck, 1991b,a, 2006, Depraetere, 1995b,a). While telicity evokes the potential actualization of a situation out of a discursive context, boundedness represents the actual realisation of the situation in the context. Eventualities are telic or atelic, and they can be realized contextually as *bounded* or *unbounded*. For example, *running a mile* is a telic situation. It can be expressed in an utterance as bounded as (3) or unbounded as (4). Boundedness is sensitive to the context of the utterance.

(3) Max ran the one-mile race.

(4) Max is running the one-mile race.

Depraetere (1995b, p. 2-3) comments that "(a)telicity has to do with whether or not a situation is described as having an inherent or intended endpoint; (un)boundedness relates to whether or not a situation is described as having reached a temporary boundary". Bounded eventualities are situations perceived by language users (i.e. the actual realization of a situation) as having reached a temporal boundary, irrespective of whether the situation has an intended or inherent endpoint.

[4]A stative situation is conceived as taking place or being done, it is unchanging and therefore homogeneous throughout its duration (i.e., it does not include stages). Situations that are not static are called dynamic situations. Such a situation may be punctual (momentary) or durative. A situation is agentive if it is caused/performed/instigated by an agent. States are by definition non-agentive.

Unbounded eventualities on the contrary are perceived as not having reached a temporal boundary. However, as already mentioned, boundedness can change, and unbounded situations may become bounded contextually through linguistic markers such as tense, grammatical aspect, noun phrases, prepositional phrases and temporal adverbials. An atelic eventuality such as *leak* may be expressed as an unbounded atelic situation as in (5), as an unbounded telic by changing the NP as in (6), or it can be turned into an telic bounded situation as in (7) through the perfective aspect (examples taken from Depraetere (1995b, p. 9)).

(5) Petrol was leaking out of the tank.

(6) The petrol was leaking out of the tank.

(7) The petrol leaked out of the tank.

There are several linguistic tests that may be used for judging an eventuality as bounded or unbounded, such as:

- The first is the compatibility with *in* or *for* temporal adverbials.
- The second is the *homogeneity* ontological feature, that refers to situations which have internal phases or stages, each of which is considered as being slightly different from the previous stage.
- The third is the entailment with the progressive, namely, if one stops while V+ing, one cannot say one has V-+ed (V stands for verb).

According to these tests, unbounded situations are homogeneous (generally states and activities) co-occur with *for* adverbials and pass the entailment with the progressive test. For example, the eventualities referred to in (5) and in (6) are homogeneous, they may co-occur with *for hours* adverbial, and they entail that the petrol *has leaked* if the event is interrupted. On the contrary, bounded situations are non-homogeneous (generally accomplishments and achievements), co-occur with *in* adverbials and do not pass the entailment with the progressive test. For example, the eventuality referred to in (7) is non-homogeneous, may co-occur with *in two hours* adverbial and does not entail that the petrol *has leaked* (in the sense that all of the petrol has leaked) if the event is interrupted.

In this research, human annotation experiments were organized in order to identify the bounded or unbounded status of each SP occurrence, as described in Section 3.

2.2 Machine Translation

Broadly speaking, a common phrase-based SMT system is the product of the combination of several components, none of which involves

linguistic knowledge. First, a phrase translation model which, trained on aligned (both at the sentence and word levels) parallel corpora, computes probabilities of translation for all sequences of words in the source text. Second, a language model, which estimates how much a candidate translation conforms to fluent target language. Third, the reordering model which predicts the changes in word order between the two languages. In order to produce a translation these components are combined during the decoding process. Here a decoding algorithm combines the translation options, making several hypothesis translations, and ultimately chooses the best one according to the language model and the reordering model (Koehn, 2010). This process is depicted in Figure 2.

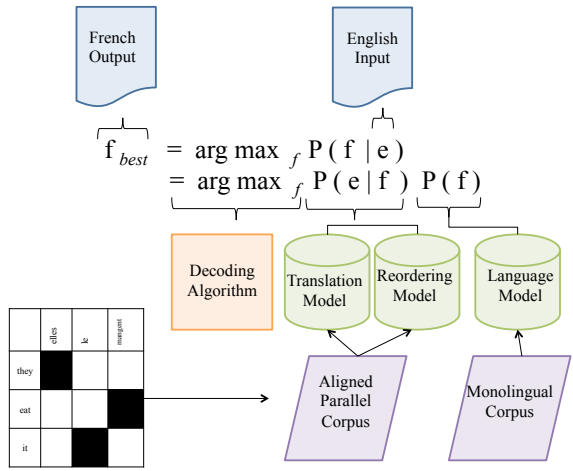

FIGURE 2 General SMT system architecture.

This architecture has proven efficient and very successful. Indeed, in the last years, SMT methods have made great progress in terms of translation output. The translation between English and French in particular has been actively used for the development of new algorithms and reached one of the best quality baselines across different language pairs (Bojar et al., 2014).

Using this type of system, however, any linguistics knowledge must be explicitly added as an additional component of the general pipeline. The translation of verb tenses is an example of one of the linguistic issues which has been addressed in this manner, as shown in the liter-

ature cited below. Indeed, verb tenses contribute to a text's cohesion and coherence, as Ye et al. (2007, 521) assert,

> Correct translation of tense and aspect is crucial for translation quality, not only because tense and aspect are relevant for all sentences, but also because temporal information is essential for a correct delivery of the meaning of a sentence.

2.3 Automatic Classification of Verbal Tenses

As mentioned before, languages differ in their encoding and usage of the tense and aspect categories. This produces mismatches, for instance, when translating VPs into a morphologically rich language from a less rich one or between morphologically poor languages. The studies described by Ye et al. (2006) and Ye et al. (2007) are motivated by this type of temporal divergences in automatically translated texts from English to Chinese. Two issues are problematic for this pair of languages. Firstly, English encodes tense whereas aspect is implicit (with the exception of the progressive marking). Secondly, Chinese does not encode tense and aspect grammatically. When Chinese aspect is marked, it takes the form of a separate word, i.e. the *le* marker, which aligns poorly with English tensed verbs, and so the aspectual information is dropped. As a result, instead of producing (8)[5] SMT systems produce the sentence (9), using the infinitive form of the verb and, in this case, with a different lexical choice.

(8) Wo ji le yi feng xin gei ta.
1st send PERF one QUA letter PP 3rd
'I *sent* him a letter.'

(9) Wo xie yi feng xin gei ta.
1st write one QUA letter PP 3rd
'I *write* him a letter.'

These studies propose a classification task of verb tense to address this problem.[6] The first study (Ye et al., 2006) tests if latent features, inspired by how humans interpret temporal relations in text, are better than surface or syntactic features for predicting the English tense of Chinese verbs. Three tenses are considered: present, past and future.

[5] Legend: 1st - first personal pronoun, 3rd - third personal pronouns, PERF - particle of completed and perfective eventuality, PP - preposition (*to/for/for the benfit of*, QUA - quantifier

[6] A classification task is a learning scheme in which an automatic classifier "is presented with a set of classified examples from which it is expected to learn a way of classifying unseen examples." (Witten et al., 2011, 40).

They found that a classifier trained just on surface features reaches 75.8% accuracy, while a classifier trained on latent features reaches 80% accuracy. The best results, 83.4%, are obtained when both types of features are combined. Latent features included telicity, punctuality and temporal ordering between adjacent events, which are also reported to be the best features. The authors argue, consequently, in favor of using latent features for tense prediction.

In the second study, their objective is to predict the appropriate Chinese aspect marker and to insert it in the Chinese translation. A classifier is trained on 2,723 verbs annotated with one of four possible Chinese aspect markers. They obtained a general accuracy of 77.25%. Contrary to the previous study, however, a feature utility ranking showed a low impact of the aspectual features of punctuality and telicity.

2.4 Machine Translation of Verbal Tenses

In these previous studies the classification results were not embedded in a SMT system and the classifier classes were the actual verbal tenses. Besides, the classifier's classes were actual verbal tenses. In contrast, Meyer et al. (2013) use classification as a means of enhancing a SMT system with knowledge about *narrativity* in order to produce better tense choices at translation time. Narrativity is a pragmatic property triggered by the category of tense and refers to determining the status of the temporal relations holding among eventualities. Two cases are possible: *narrative* and *non-narrative* usages of a verbal tense. A narrative usage points to the case when the two eventualities are temporally linked (both forward and backward temporal inferences). Non-narratives usages point to the case when eventualities are either not temporally linked or they occur simultaneously.

In their paper, Meyer et al. (2013) built a classifier, which was trained on a small manually-annotated corpus with narrativity, to generate narrative and non-narrative disambiguation labels for the English SP verbs of a large parallel corpus. In other words, they classify the SP verbs of the SMT training data into narrative or non-narrative instances. With this second corpus, they built a SMT system using a factored model of translation (explained in section 5). This system gained 0.2 BLEU points[7] when compared to a baseline system lacking the disambiguation labels. The authors note two shortcomings in their method. The classification results are rather moderate (F1 = 0.71), since narrativity is hard to infer from surface clues. Furthermore, they note a problem

[7]The BLEU score is an automatic measure of precision computed on the comparison between the translation produced by the system and a human reference translation (Papineni et al., 2002).

with the identification of the SP verbs in the large corpus, in particular when used in the passive voice (for instance, instead of "was taken", they only detect "was"). Following Meyer et al. (2013), in our work we built a classifier trained on a small manually-annotated corpus and then used the classifier to annotate a large corpus for training a SMT system. In our study, the corpus annotation concerns *boundedness*. Each SP instance is annotated with a *bounded* or *unbounded* label and these labels are then used as disambiguation markers. Compared to *narrativity*, *boundedness* is more likely to be correctly learned by a classifier on the basis of surface clues and linguistically-informed features. Finally, we use a more sensitive method to identify English SP verbs either in the active or passive voice.

In comparable work, Loáiciga et al. (2014) automatically identify all English and French VPs in a large parallel and aligned corpus. Next, they automatically annotate the VPs in each side of the corpus with one of 12 tenses indicating present, future or past time. The annotation allowed them to map and to measure the distribution of tense translation between the languages. They find that the ambiguity of the translation of English SP into French PC, IMP, PS and PRES is significant. Using this automatically annotated corpus, the authors present two SMT experiments for disambiguating the translation of the English SP into French.

Firstly, the parallel and aligned corpus is used to automatically annotate the English verb with the French tense. For instance if the verb *ran* is translated as *courait* an *imparfait* label is used, if a second instance of the same verb is translated as *a couru*, then a *passé composé* label is used. They train a SMT system on this annotated corpus and obtain an increase of 0.5 BLEU points over a baseline with no French tense labels. This experiment is intended as an oracle measure of how much improvement one could expect if the system knew all French tenses before translation. In a second experiment, the authors use the corpus to train a classifier of French tense translation using features from the English side only. In other words, the gold French tense annotation is not used, instead tense labels are predicted. The classification task is not trivial since it involves nine classes (nine tenses)[8] inferred from the source language. Results vary significantly depending on the particular tense. Finally, they build a second SMT system using the French tense labels produced by the classifier and, hence, error prone. This second system performance increased by 0.12 BLEU points over the baseline.

[8]Only a subset of the 12 annotated tenses were considered in the classification experiments of this work.

They note that the quality of the translation was determined to a great extent by the quality of the classifier for each particular tense.

Meyer et al. (2013) and Loáiciga et al. (2014) present the only existing work on machine translation of verbal tenses between English and French. Otherwise, most of the work on machine translation of verbs concern the translation between Chinese and English. Indeed, the grammatical aspect markers for *perfective* and *imperfective* are optional in Chinese. Therefore, Chinese verbs are underspecified when compared to English, and what in English would correspond to present and past tenses, for example, are hard to distinguish in Chinese, compromising the quality of translation. Addressing this problematic, Olsen et al. (2001) report probably the work most closely related to our own. The particular architecture of their system (interlingua model) allows them to obtain reliable lexical information associated with each verb. This information includes primitives (GO, BE, STAY, ...), types (Event, State, Path, ...) and fields (locational, temporal, possessional, identificational, perceptual, ...). Using this information and some heuristics which exploit additional clues from the sentence such as adverbs, they implement an algorithm that identifies *telic* Chinese verbs. Their hypothesis is that Chinese sentences with a telic aspect will translate into English past tense and those without the telic aspect as present tense.

Their system is tested on a 72 verb test set matched against a human reference translation. Results are given in terms of accuracy or correct translations. While the baseline system obtained 57% correct translations, a second system which uses the telic information of verbs obtains 76% correct translations. Furthermore, a third system built using the telic information along with other linguistic information such as grammatical aspect and adverbials obtained 92% accuracy. Contrary to our framework, this system is highly deterministic, with a fixed correspondence *+telic* → *past*, *−telic*→ *present* which might be incorrect in other language contexts. Besides, the identification process of telic verbs relies heavily on their particular system's lexicon, making it difficult to implement in different systems.

In the same context of Chinese to English translation, Gong et al. (2012b) propose a method to reduce tense inconsistency errors in MT. First, they determine the tense of the main verb for each sentence on the English side of a parallel corpus based on heuristics and POS-tags. The sequence of all tenses in the sentence is defined as "intra-tense", while the tense of the main verb of the sentence is defined as "inter-tense". For example, given a sentence, a sequence like {present*, present} is its intra-tense, where * represents the main clause tense or its inter-tense. Using this type of sequences, they compute n-gram statistics

and probabilities to build a tense-model out of the English side of the corpus.

At decoding time, when a hypothesis has covered all source words, the intra-tense of the current sentence is computed and then the inter-tense of the previous sentence. With this information, the hypothesis is re-scored, including the weight of each tense-feature (inter and intra) using MERT (Och, 2003). They gain 0.57 BLEU points using the inter-tense; 0.31, using the intra-tense; and, 0.62 using the combination of both.

The same authors report on a follow-up study (Gong et al., 2012a) which additionally uses information concerning the source language Chinese to extract the features given to the classifier. This classifier is trained to assign one of four tense labels to Chinese verbs before translation. Each of these labels has an a associated probability, and the highest one is retained. As before, during decoding time, this probability is fed to the SMT system and the hypothesis translations are re-ranked. They obtain a BLEU score improvement of 0.74 points.

Finally, as part of a study mostly interested in reordering English verbs when translating into German, Gojun and Fraser (2012) report a pilot experiment concerning verb tense disambiguation. They trained a phrase-based SMT system using POS-tags as disambiguation labels concatenated to English verbs which corresponded to different forms of the same German verb. For example the English *said* can be translated in German using a past participle *gesagt* or a simple past *sagte*. This system gained up to 0.09 BLEU points over a system lacking the POS-tags.

3 Human Annotation

The corpus data used in our work was provided by Grisot and Cartoni (2012). It is the same corpus as that used in Meyer et al. (2013)'s study. Grisot and Cartoni (2012) built a parallel corpus, consisting of texts originally written in English and their translation into French. These texts were randomly selected and belong to the following stylistic registers: literature, journalism, discussions of the European Union Parliament (the Europarl corpus) and European Union legislation (the JRC-Acquis Corpus). Grisot and Cartoni (2012) found that English SP was not translated by a single or specific FR tense. In order to investigate this translation divergence, Grisot and Cartoni built a smaller sub-corpus with 435 English sentences containing SP tokens and their corresponding French translation. This corpus was manually-annotated with linguistic (semantic and pragmatic) properties, as suggested in

Grisot (2015): *narrativity*, *perfectivity* and *boundedness*. The human annotation for each of these features was independent of the others. For each feature, the judges had access only to the English data containing SP occurrences and to the annotation guidelines. The annotated data was analyzed in its totality after the various human annotation were done (Grisot, 2015).

Meyer et al. (2013) made use of this corpus annotated with *narrativity* in their MT experiments. Here, we use the same corpus and focus on *boundedness* and its utility for determining the verbal tense used in a target language. In what follows, we will first describe the human-annotation experiments with *boundedness*.

3.1 Participants, Procedure and Materials

Two human judges participated at the annotation. One was one of the authors of this study, graduate student at the time. The second human judge was a postdoctoral research peer fluent in written and spoken English. The participation at the experiment was not paid. The two judges received annotation guidelines, which contained the definition of boundedness, examples illustrating bounded and unbounded situations and the three tests presented in section 2.1. These linguistic tests can be summarised as follows. Bounded eventualities take *in* adverbials, are not homogenous situations and do not pass the entailment with progressive test. Unbounded eventualities take *for* adverbials, are homogenous situations and pass the entailment with progressive test. For instance, in the first example below, the VP *wrote an email* was judged as *bounded* because: it takes *in* adverbials such as *He wrote an email in five minutes*, it does have internal phases and, if one stops writing an email, one has not written the email. On the contrary, in the second example below, the VP *sat behind a huge desk* was judged as *unbounded* because: it takes *for* adverbials such as *He sat behind a huge desk for two hours*, it does not have internal phases and, if one stops sitting behind a huge desk, one has sat behind a huge desk.

(10) John entered in the president's office. The president *wrote an email*.

(11) John entered the president's office. The president *sat behind a huge desk*.

For each sentence, the two participants were asked to judge in the context the eventuality referred to by the verb in italics according to the three linguistic tests provided. The results of the annotation experiment were analysed from two perspectives. Firstly, in a monolingual perspective, the Cohen's κ coefficient (Carletta, 1996) was used to measure the

inter-annotator agreement rate. Secondly, the labeled items were compared to a reference baseline containing the verbal tenses used for the translation of the SP in the French part of the parallel corpus.

3.2 Results

Judges agreed on the label for 401 items (92%) and disagreed on 34 items (8%). The agreement rate corresponds to a κ value of 0.84. All 34 disagreements were resolved in a second phase consisting of a discussion between the two judges, corresponding to a κ value of 1. The κ values of both phases of annotation indicate that the judges understood the annotation guidelines and that their judgments are reliable. The data contains 236 SP tokens judged as *bounded* and 199 as *unbounded*, that is 54% and 46% respectively. The data containing agreements from both annotation rounds (435 items) was investigated cross-linguistically by looking at the verbal tenses used in the French parallel text.

Most frequently, bounded eventualities correspond to a translation with a PC or PS and unbounded eventualities correspond to a translation with an IMP for 360 items (82%) as illustrated by the first two examples in Figure 4. Using a chi-square test for independence, this correlation is shown to be statistically significant ($\chi = 182.62$, df=1, $p < .001$). Figure 3 depicts the relationship. The less frequent cases, namely bounded eventualities corresponding to a translation with an IMP and unbounded eventualities corresponding to a translation with a PC or PS are illustrated in the last two examples in Figure 4. This lack of perfect one-to-one correspondence points in favour of a non-deterministic MT system and discourages rigid constraints of the type *bounded* $\rightarrow$ PC, *unbounded* $\rightarrow$ IMP.

The corpus data used in this experiment contains further manually annotated information, such as grammatical aspect (coming from another human annotation experiment), narrativity (coming from yet another human annotation experiment), verbal tense used in French and the infinitive form of the verb, as described by Grisot (2015). Figure 4 presents an example of the corpus data. In the first example, the simple past form *asked* was annotated with the perfective aspect, was judged as having a narrative usage in this context (i.e. the eventuality *reveal* temporally precedes the eventuality *ask*), was judged as being bounded (*reveal* is a punctual situation - achievement- it combines with *in* adverbials, it is not homogeneous and it does not pass the entailment with the progressive test) and its translation into French in the parallel corpus is in the PC. In the second example, the simple past form *fascinated* was annotated with the imperfective grammatical aspect, was judged as having a non-narrative usage (the eventuality *fascinate* is

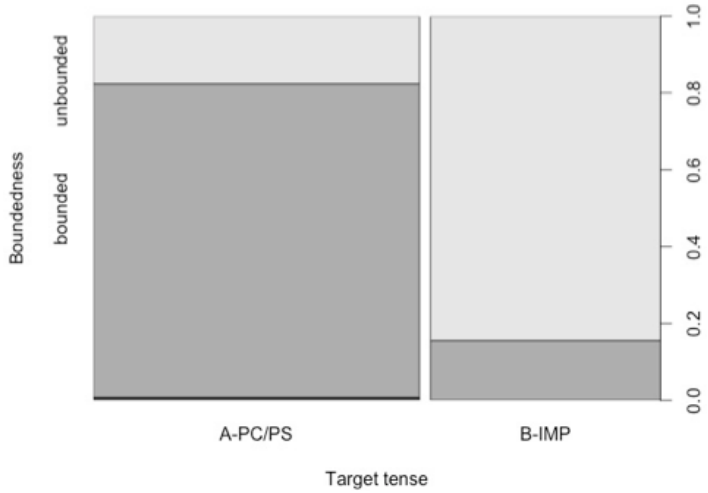

FIGURE 3 Relationship between lexical aspect and verbal tense

temporally simultaneous with the eventualities *be worth* and *be wilful*), was judged as being unbounded (*fascinate* is an activity, it combines with *for* adverbials, it is homogeneous and it passes the entailment with the progressive test) and its translation into French in the parallel corpus is in the IMP. A similar analysis can be provided for the last two examples from Figure 4.

The 435 SP tokens annotated with boundedness among others types of information were used as an empirical basis to develop a linguistic model through inferential statistics. They are however not sufficient to provide meaningful knowledge to a SMT system, which needs much larger quantities of data. For this reason, as described in the next section, the manually annotated corpus is used as training data to build a classifier which labels new data with information about aspect. The classifier approach permits the assessment of the manual annotation and the creation of large quantities of annotated data to build a SMT system. The following sections are dedicated to describing, on the one hand, experiments on the prediction of *bounded* and *unbounded* labels for English SP verbs, and on the other hand, experiments with a SMT system enhanced with the labels mentioned.

4 Predicting the *boundedness* of English SP Verbs

Investigating the pragmatics of temporal reference, Grisot (2015) proposed that it may be expressed through several linguistic expression

Sentence	Verb	Grammatical Aspect	Narrativity	Bounded-ness	FR tense	Infinitive
In one instance, Kazakhstan revealed the existence of a ton of highly enriched uranium and *asked* the United States to remove it, lest it fall into the wrong hands.	asked	perfective	narrative	bounded	PC	to ask
He *fascinated* everybody who was worth fascinating and a great number of people who were not. He was often wilful and petulant, and I used to think him dreadfully insincere.	fascinated	imperfective	non-narrative	unbounded	IMP	to fascinate
A few days ago, in a manner of speaking, we *said* that Bin Laden had provided the impetus for implementing methods for fighting terrorism that the Commission had been planning and that Parliament had requested some time ago.	said	perfective	narrative	bounded	IMP	to say
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.	were seen	imperfective	non-narrative	unbounded	PC	to see

FIGURE 4 Example of human annotated corpus data

markers, such as verbal tenses (expressing both tense and grammatical aspect), lexical aspect, temporal connectives and adverbials, as well as non-linguistic types of information, such as contextual and world knowledge. Linguistic markers are correlated and the value of a marker can be predicted based on the values of the other markers. The leading hypothesis in the experiments presented below is that the linguistic information available in the sentence where the verb occurs can be used to predict the boundedness status of English SP verbs due to its context-dependent character.

In this section, our motivation is threefold. Our foremost goal is to predict the the boundedness value of English SP verbs from the corpus previously described. The second goal is to understand better the role of each of the linguistic factors in predicting this pragmatic component of the lexical aspect of SP verbs, by isolating each of them. Our third and final goal is to propose a classifier trained on automatically generated features only. Since human annotated data is expensive and time consuming, training a classifier for the task would expedite the annotation of large amounts of text.

4.1 Experiment 1

In this first experiment, a classifier is trained for predicting the type of *boundedness* of the English SP verbs contained in the corpus presented

in Section 3. The additional linguistic annotations of the corpus are exploited as features for the classifier. Additional syntactic and temporal features automatically generated and extracted from the sentence in which the verb occurs are also included. Since this is a fully supervised classifier partially fed with features known to be pertinent for the task, its results are expected to be a measure of the maximum success rate on this particular task.

Data and Tools

The Stanford Maximum Entropy package (Manning and Klein, 2003) is used to build a Maximum Entropy classifier. This classifier is roughly based on multiclass logistic regression models and it is an appropriate model in a context such as ours, where the number of training examples is limited relative to the large number of features we generate. The corpus described in Section 3 is used both as training and testing data. Given its small size, results are reported as averages over ten-fold cross-validation for the two experiments which follow. Note that the ten-fold validation procedure ensures lower variance and maximum generalization power given that our corpus is very small. *Boundedness* is the prediction class and it has two possible values *bounded* or *unbounded*.

Features

The features used are of two types: syntactic and temporal. Syntactic features model the context (i.e. the sentence) in which the English SP verb occurs, whereas temporal features refer to the interpretation or meaning of the SP verb itself. Manually annotated features which were taken from the previous corpus annotation scheme are indicated by a * symbol. For the automatically generated features, the dependency parser of Bohnet et al. (2013) from MateTools was used on the English side of the corpus to produce part-of-speech tags and dependencies labels.

Syntactic features

1. Simple past verb token*: this refers to the English SP verb to be classified.
2. Infinitive form*: the non-finite form of the English SP verb token. Since the lexical aspect is intrinsic to the verb form, we consider pertinent to use the base form.
3. Grammatical aspect*: a pragmatic feature taking the values of *perfective*, which stresses the initial and final boundaries of an eventuality, or *imperfective*, which does not stress these boundaries.
4. French tense*: the tense of the French translation corresponding

to the English SP verb in the parallel corpus.

5. Position in the sentence: refers to the ordinal position of the English SP verb in the sentence.
6. POS-tags of the English SP token: they distinguish between active voice SP verbs, e.g., *went* (VBD); compound active voice SP verbs e.g., *did go* (VBD+VB); and passive voice SP verbs, e.g., *was taken* (VBD+VBN).
7. Head and its type: it refers to the syntactic head of the verb to classify, along with its POS-tag.
8. Children dependencies: they indicate the dependency relation of the three nearest children of the English SP verb.
9. Children POS-tags: they indicate the POS-tags of the three nearest children of the verb. With this and the previous feature, we expect to capture some of the linguistic reflexes of aspect (Section 3), for example the presence of *in* prepositional phrases for *bounded* eventualities.

Temporal features

10. Adverbs: Meyer et al. (2013) manually gathered a list of 66 adverbial (temporal) expressions; we checked for the presence or not of such expressions in the English sentence.
11. The type of adverb: additionally, each adverbial expression was labeled by Meyer et al. (2013) as a marker of synchrony (e.g., *meanwhile*) or asynchrony (e.g., *until*). These type labels were also included among the features.
12. *Narrativity**: a pragmatic feature referring to the temporal structure between eventualities. It can have the values of *narrative* or *non-narrative*.

Results

	Bounded	**Unbounded**
Precision	0.8833	0.8909
Recall	0.9038	0.8650
F1	0.8943	0.8759
Accuracy	0.8857	

TABLE 1 Average classification results of Experiment 1 using ten-fold cross-validation

Results show a very good performance of the classifier, reaching up to 0.8943 F-score for the bounded class and 0.8759 for the unbounded class. These results are partially explained by the features taken from

the previous annotation of the corpus, produced by expert linguists. However, even if all features are pertinent and linguistically-motivated, they are not error-free. Those generated using an automatic tool in particular may introduce some noise, although the general performance of the parser used is very good. In what concerns the gold annotation of the *bounded* and *unbounded* labels, they contain some degree of ambiguity as well. As expressed by annotators, judgments can be ambiguous since they also depend on the particular context each verb appears in. We therefore think that these results reflect to some extent the intrinsic ambiguity of the *boundedness* of English SP verbs.

4.2 Experiment 2

The main goal of this paper is to enhance a SMT system with *boundedness* as a means to disambiguate English SP verbs when translating into French IMP, PRES, PC or PS. For building a SMT system, a 435 sentences corpus is clearly insufficient, a much larger parallel corpus is needed. As in the previous experiment, here a classifier is trained for predicting one of the two values for *boundedness* of the English SP verbs. However, the objective of this second classifier is to approximate the results obtained in Experiment 1, using a sub-set of the features previously described before in 1 to 12. This sub-set is composed of those features which it is possible to generate from raw data. Consequently, the results of this experiment are expected to give a realistic impression of the quality of the *boundedness* detection task on a large corpus using automatically generated features and a small quantity of annotated data (the only annotation being the gold prediction class) for training.

Data and Tools

As in Experiment 1, a Maximum Entropy classifier is built using the Stanford Maximum Entropy package (Manning and Klein, 2003). The dependency parser of Bohnet et al. (2013) from MateTools is used on the English side of the corpus. Additionally in this experiment, the TreeTagger tagger (Schmid, 1994), which produces POS-tags and lemmas for all words in the sentence, is used on the English side of the corpus as well. The corpus described in Section 3 is used as training and as testing data. Results are reported as averages over 10-fold cross-validation. As before, *Boundedness* is the prediction class and it has two possible values *bounded* or *unbounded*.

Features

In Experiment 1, the manual annotation already existing in the corpus was recovered as features for the classifier since it was known to be pertinent for the task. Some of those features can be easily recreated using

syntactic and morphological parsers. However, this is not the case for *grammatical aspect*, *French tense* and *narrativity*. In this second Experiment, the input to the classifier is limited to the features which will be available when using the parallel SMT data, those created automatically. Following the same intuition as before, the training features are divided into syntactic and temporal types.

Syntactic features

1. Simple past verb token: this refers to the English SP verb to be classified. In this experiment, we used the heuristics based on POS-tags described by Loáiciga et al. (2014) to identify all English SP instances in the sentence.
2. Infinitive form: the non-finite form of the English SP verb. It was extracted from the output produced by TreeTagger tagger (Schmid, 1994).
3. Position in the sentence
4. POS-tags of English SP token
5. Head and its type
6. Children dependencies
7. Children POS-tags

Temporal features

8. Temporal adverbs
9. The type of adverb

Results

	Bounded	**Unbounded**
Precision	0.8142	0.8509
Recall	0.8747	0.7578
F1	0.8401	0.7944
Accuracy	0.8224	

TABLE 2 Average classification results of Experiment 2 using ten-fold cross-validation

Table 2 shows the results. Note that in the first experiment, one SP verb per sentence is annotated. In this experiment we identified all English SP instances in a sentence automatically. Nonetheless, results reported in Table 2 are limited to the same verbs annotated originally and used in Experiment 1. Hence, results of both experiments are comparable. For the subsequent SMT experiments, all English SP verbs are identified and tagged as *bounded* or *unbounded*.

The quality of the classifier is quite satisfactory, reaching up to 0.8401 F-score for the bounded class. Results are reasonably comparable to those of Experiment 1. In this experiment, however, the unbounded class seems a bit harder to predict than in Experiment 1, as evidenced by the generally lower figures, recall in particular.

4.3 Discussion

Experiment 1 showed that *boundedness* could be accurately predicted from sentence features. These features were partially annotated by hand and they were expected to be relevant for the task. Experiment 2 produced good quality results despite the partially missing gold information used in Experiment 1 (i.e., grammatical aspect, French tense, narrativity). While Experiment 1 set the upper bound of the task, the results of Experiment 2 were established under more realistic conditions, since automatic tools were used to generate the features (which implies some noise). The second experiment measured the quality with which completely raw data can be automatically annotated. As expected, results of Experiment 1 are better than those of Experiment 2, since only a limited set of the features was used in the second experiment. There is a significant difference of about 8% in performance between the two classifiers ($\tau(434) = 7.28$, p-value = 1.5e-12). Yet, the second classifier was still able to learn how to discriminate between *bounded* and *unbounded* SP verbs in a satisfactory manner.

To measure the impact of the result further, we set a baseline based on randomisation for comparison. A random sample with resampling of 435 *bounded/unbounded* labels with probabilities 0.54 and 0.46 respectively was generated. These probabilities correspond to the distribution of the labels in the human annotated corpus (Section 3). Next, we compared the obtained random labels to the gold corpus in order to compute precision, recall and F-score in the standard fashion (Table 3). Both the results of Experiment 1 and Experiment 2 are significantly better than our random sample ($\tau(434) = -76.71$, p-value = 2.2e-16; $\tau(434) = -57.05$, p-value = 2.2e-16), which further indicates that the prediction results are solid. A graphical summary of this comparison is given in Figure 5.

To judge the predictive power of each of the features involved, feature ablation for each of the experiments was done. We compared the performance of the classifier trained on all the features to its performance when each feature is subtracted (one at the time) from the model. For each feature removal round, we used ten-fold cross validation and calculated the F-score for each class. The observed changes are plotted in Figures 6 and 7; the mean of all the folds is given by the thick middle

	Bounded	**Unbounded**
Precision	0.5574	0.5192
Recall	0.5763	0.4426
F1	0.5667	0.4779
Accuracy	0.5402	

TABLE 3 Results of a sample of 435 randomly generated labels according to their gold distribution probability.

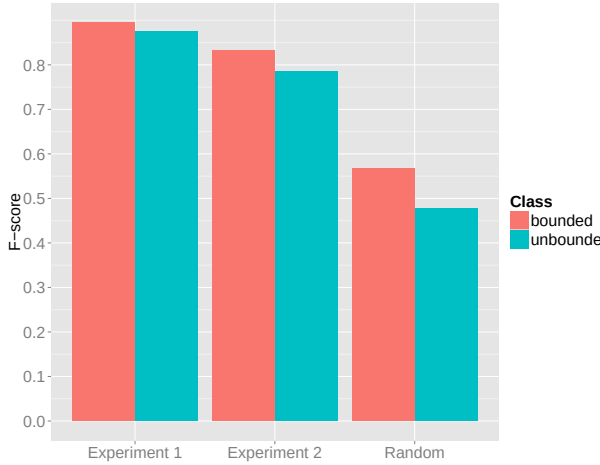

FIGURE 5 Comparison of results obtained in the classification Experiments. The blue represents the *bounded* class and the red represents the *unbounded* class.

line in each boxplot.

In both experiments, the interaction of the features in dependent on the class to be predicted. For example, grammatical aspect and narrativity seem to be important for the unbounded class only, while the verb's POS tags seem to be more informative for the bounded class. However, it is clear that the adverbs and the infinitives are the features with the most predictive power for both classes and in both experiments. Interestingly, the French tense does not contribute as much to the model as expected. Moreover, although we initially thought that the verb position could be an indicator of main (lower values) vs subordinated verb status (higher values), the analysis of the results indicated that it is not very informative. The verb's children dependencies are another feature which did not provide improvements to the model. The

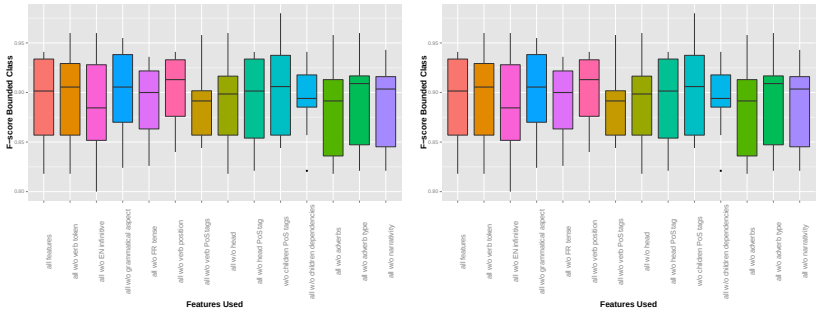

FIGURE 6 Feature ablation comparison for Experiment 1.

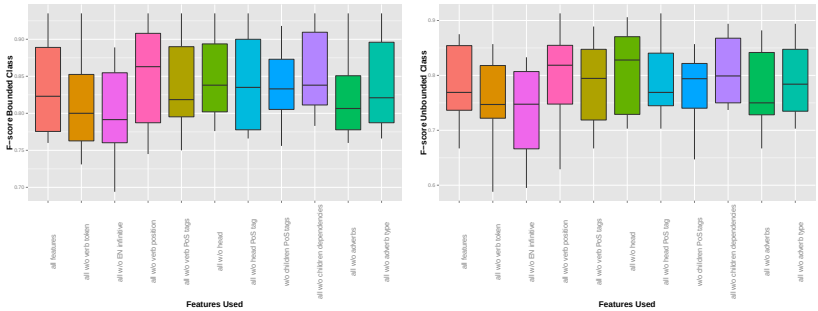

FIGURE 7 Feature ablation comparison for Experiment 2.

children POS tags is a more useful feature.

Following Meyer et al. (2013), we also experimented adding information concerning the previous verb to the feature vector of the current verb. The intuition was that modeling more context would benefit the classifier. We noted, however, a decrease in accuracy of around 15%, therefore this information was dropped. This outcome is in line with what is reported by Ye et al. (2007, 6) on aspect prediction for Chinese: "We expanded the features vector of each verb by including the features from the previous and the following verb; the results showed no improvement by adding these contextual features." In our case, this might be due to data scarcity, given the small size of the corpus.

To conclude, the results and analysis presented in this section indicate that the classifier from Experiment 2 is satisfactory and can be used to annotate large quantities of raw data. Since this classifier is trained on features created automatically only, it serves our purpose of annotating training data for building a SMT system enhanced with

lexical aspect information. The time and cost of annotating such data manually are high and often unaffordable. Being able to create good quality data automatically is valuable and highly appreciated within the NLP community.

In the next section, the classifier from Experiment 2 is used to annotate all English SP verbs in a large corpus with *bounded* and *unbounded* labels in order to train a SMT system. The purpose of the SMT experiment is to show that knowledge about boundedness is relevant for the disambiguation of the English SP verb when translated into French, since in this particular pair of languages, there is a one-to-many translation mapping. Boundedness labels should be used for improving the choice of the verbal tense form by the SMT system.

5 Using the Predictions for Machine Translation

In this final part, we assess how much a SMT system enhanced with *boundedness* knowledge improves the translation of the English SP into French. Phrase-based SMT systems often generate only the most frequent translation possibility, the PC, as the default. Our goal is to provide a system with *bounded/unbounded* labels in order to boost the other three tenses, improving the tense translation choice.

Given that there is no sufficiently large data set annotated with aspect information to train a SMT system, we annotate our own using the classifier trained before, in Experiment 2 (Section 4). The data is taken from the MultiUN corpus, a corpus of translated documents from the United Nations, provided by Eisele and Chen (2010). All English SP verbs are identified and labeled as either *bounded* or *unbounded* automatically. Table 4 shows the number of English SP verbs annotated with this method.

	Sentences	**SP verbs**
Training	350,000	134,421
Tuning	3,000	1,058
Testing	2,500	1,275

TABLE 4 Data setup of the SMT system.

We use the Moses Toolkit (Koehn et al., 2007) to build two systems: a baseline without *boundedness* labels and an aspect-aware system with the labels. Both systems are phrase-based models with an identical composition, according to the set-up presented in Table 4, and use a 3-gram language model built using KenLM (Heafield, 2011). The language model is trained over 10 million sentences of French monolingual

data taken from the 2015 Workshop on Machine Translation (WMT15) (Bojar et al., 2015). Optimization weights were fixed using Minimum Error Rate Training (MERT) (Och, 2003).

The *boundedness* labels are combined with the SMT system using a factored model (Koehn and Hoang, 2007). A factored model is a variant of the phrase-based model which integrates linguistic markup (so called factors) at the word level. In practical terms, this means that instead of the standard text, the system is trained on annotated text of the form shown in (12). In our system, only one factor, i.e. the bounded or unbounded label is used. Accordingly, the SP verb *ran* in sentence (12) has *unbounded* as its factor, while all the other words have NULL as default factor. For verbs composed of multiple words, (e.g., *cut off, was made, were laid down*) all words are labeled with the same *bounded* or *unbounded* factor.

(12) Max ran for an hour.
Max|NULL ran|UNBOUNDED for|NULL an|NULL hour|NULL .|NULL

In the model, factors are taken into account in a non-deterministic manner. In other words, there is no exact mapping between a given label and a particular output. For instance, a *bounded* label does not necessarily lead to a verb with the PC French verbal tense. Instead, factors are considered when estimating the translation probabilities computed over the entire parallel corpus.

The results obtained are given in Table 5 using the BLEU (Papineni et al., 2002) score. The BLEU score computes the matches between the output of the system and a human reference translation at the sentence level and is the de-facto metric used in the MT domain. Its numerous flaws are well-known, e.g., it performs less well with a single reference, does not provide any clue on qualitative criteria such as lexical choice, and it does not account well for recall (Song et al., 2013). Nevertheless, an increase in BLEU is generally correlated with an improvement of overall translation quality.

In our case, the system with the *boundedness* labels (aspect-aware) obtained an increase of 0.98 BLEU points. When computing the BLEU score on the sentences with SP verbs only, we obtained a difference of 1.58 points. These scores reflect an improvement in the quality of the SP translation. On the one hand, these increments suggest that the method is not degrading the general translation quality of all the other words in the sentence; on the other hand, they suggest that it is not changing the SP translations estimated as already adequate by the baseline model. This result is non negligible, considering in particular

that the aspect-aware system targets only SP verbs and not all words in the sentence (BLEU, being an exact-matching-oriented metric, increases as the number of matching words to the reference increases).

System	BLEU test set	BLEU SP sub-set
Baseline	31.75	30.05
Aspect-aware	32.73	31.63

TABLE 5 BLEU scores of the SMT systems computed on the test set and and on the sentences with SP verbs only.

Since automatic MT scores are not very informative and can be difficult to interpret, a bootstrap resampling significance test as introduced by Koehn (2004) was carried out. This test estimates the difference in performance of one SMT system in comparison to another. Using the test set, 100 paired samples of 300 sentences each containing at least one verb in the SP tense are generated. Then a BLEU score is computed and recorded for each sentence in each sample. Results are given in Figure 8. Consistently across all the samples, $\approx 50\%$ of the sentences containing at least one English SP verb were better translated by the aspect-aware system than by the baseline system. Furthermore, following the methodology proposed by Zhang et al. (2004), a 90% confidence level estimate computed over the 100 samples places the confidence interval of the differences in BLEU scores between the two systems at [0.62, 2.85].

Automatic metrics and statistical tests do not give any further indications on the particular qualitative differences in the tense translation between the outputs. To overcome this, a manual evaluation of 200 randomly selected English SP verbs was done as well. The selection contains an even distribution of the labels. Results are summarized in Tables 6 and 7.

Table 6 shows the assessment of the classifier performance. The verb boundary identification is very good with 91% accuracy. As mentioned, verb identification was done automatically, using POS tags along with a set of heuristics. Errors are mostly due to incorrect tagging of some ambiguous cases such as the construction *was concerned* in which only *was* is identified. Another common error occurred with adjectives identified as verbs, for instance *titled* in *... under the item titled "the Situation in the Middle East"*. On the labeling, the manual evaluation shows 65% accuracy, a figure lower than those obtained automatically and presented in Table 2.

In general, the *bounded* class seems more difficult to predict than

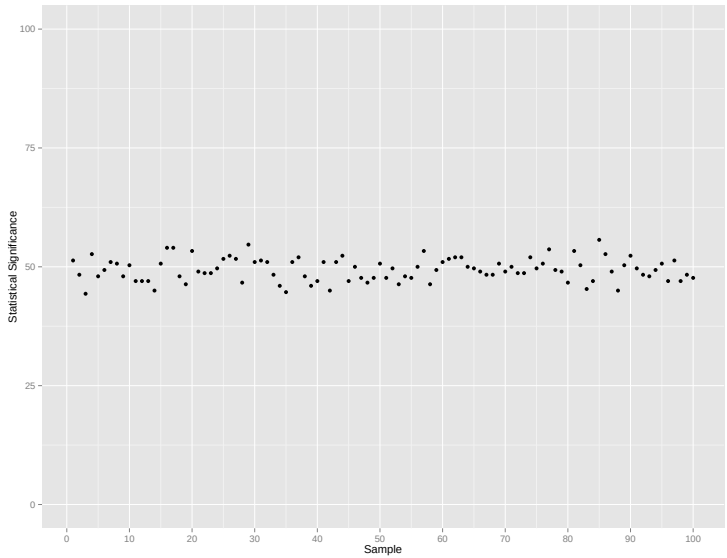

FIGURE 8 Results of paired bootstrap with resampling test. The x axis shows the ID of the samples and the y axis displays the percentage of sentences per sample which obtained a higher BLEU score than the sentences in the baseline.

the *unbounded* one. The manual evaluation also revealed that several verbs which usually express one-time events, as 'ask', 'request', 'result', 'adopt', 'add' or'call', were treated as having a duration which is much less common. Finally, we noticed that several instances of the same verb appear repeatedly, therefore, the same classification error was repeated (e.g. *was* labeled as bounded).

Table 7 presents the results of the comparison between the baseline and the aspect-aware systems. The classification of the English SP verbs was correct in 65% of the cases and their translation is improved in 25% of the cases. Most verb translations are unchanged, most probably because the weight of the *bounded/unbounded* factor yields the same best translation hypothesis as the baseline, in other words, the same translation probability would be produced without the factor. Remember that the translation distribution is highly skewed in favor of the PC. Last, 21% of the examples were degraded, a possible outcome given the non-deterministic disposition of the factored model. This result is also directly linked to the results of the *bounded/bounded* labeling: correct labels entail twice as many improved translations (31 vs 17).

	Correct	Incorrect	Total
Verb Identification	182 (91%)	18 (9%)	200
Predicted Label	117 (65%)	65 (35%)	182

TABLE 6 Manual evaluation of the classification results of 200 SP verbs. The predicted label is evaluated only on those cases where the SP is identified correctly.

Bounded/ Unbounded	**Translation**		
	Improved	**Unchanged**	**Worsened**
Correct	31	69	17
Incorrect	15	29	21
Total	46 (25%)	98 (54%)	38 (21%)

TABLE 7 Relationship between classifier results and translation quality of the 182 correctly identified SP verbs.

Two examples of the improved translations are presented in Figures 9 and 10. In Figure 9, the verb *was* is labeled as *unbounded* and translated in French using the IMP. This is the same translation used in the reference. In Figure 10, both verbs *had* and *were* are labeled as *unbounded*. In this example, however, it may be the case that the labeling has an effect only on the first one. Both verbs are translated using the PC by the baseline. The factored model, instead, produces the IMP tense (same as the reference) for the first one but not for the second. Other differences between the translation outputs of the systems are most probably due to a different ranking of the hypothesis during decoding time. Different hypothesis combination is likely to happen when translating long sentences.

Source	Education **was** mandatory up to the age of 16.
Baseline	L'éducation **est** obligatoire jusqu'à l'âge de 16 ans.
Aspect-aware	L'éducation **était** obligatoire jusqu'à l'âge de 16.
Reference	L'enseignement **était** obligatoire jusqu'à l'âge de 16 ans.

FIGURE 9 Example outputs of the SMT systems.

The method developed in this paper has been used in relatively recent pieces of work such as Meyer et al. (2013), Loáiciga et al. (2014) and Meyer (2014). In all these papers, different types of linguistic information are used to train classifiers which generate large quantities of annotated data in order to enhance SMT systems with linguistic knowledge. For example, Meyer (2014) tested several SMT systems with knowledge related to connectives and verbal tenses. The method devel-

Source	He also considers that he has exhausted domestic remedies with regard to release on bail, and that the remedies mentioned by the State party **had** no prospect of success and **were** not available.
Baseline	Il considère aussi qu'il a épuisé tous les recours au niveau national en ce qui concerne libération sous caution et que des procédures de recours mentionnée par l'État partie **n'ont** pas de chances d'aboutir, et n'**ont** pas **été** communiquées.
Aspect-aware	Il estime qu'il a épuisé les recours internes en ce qui concerne la libération provisoire sous caution, et que les recours mentionné par l'État partie n'**avaient** aucune perspective de succès et n'**ont** pas **été** disponible.
Reference	Il estime également avoir épuisé les recours internes quant aux demandes de mise en liberté sous caution, et que les recours mentionnés par l'Etat partie n'**avaient** aucune chance d'aboutir et n'**étaient** pas disponibles.

FIGURE 10 Example outputs of the SMT systems.

oped in this paper is also motivated linguistically. The linguistic information is combined with the SMT system through a factored model. Recently other methods have been suggested such as direct document-level translation (Hardmeier et al., 2012, 2014). This method consists in a completely different strategy of translation in which the decoding algorithm itself is modified to process the text as a whole. This type of method does not need to place additional annotations or labels in the input text as we did here. Both methods have proved their efficiency when compared to a baseline system. In future research the two methods could be compared with respect to the same linguistic phenomenon.

6 Conclusions

In this paper, we proposed a method to disambiguate and improve the SMT of English SP verbs into French. This single English tense has four possible translations in French. The method combines knowledge about a pragmatic componenent of aspect which was operationalized as the *boundedness* factor in a SMT system. This method proved to have good results with respect to the targeted verbal tenses without decreasing the quality of the translation of the surrounding words in the sentence. Indeed, manual evaluation of the translated texts showed that correctly labeled verbs with boundedness presented a better tense translation. With this work, we hope to have contributed building a more natural and coherent MT output in terms of *adecuacy* and *fluency*[9], which are

[9]The first refers to to how accurately the input meaning is conveyed in the target language, whereas the second refers to grammatical correctness and word choices

two defining desirable criteria for machine translation output.

Additionally, we built two classifiers for *boundedness*, one including a larger set of features including oracle features and the other one trained on automatic features only. The first showed that boundedness can be annotated reliably and set the upper-bound performance of the classification task. The second allowed us to label a large corpus based on a minimal and affordable quantity of manually annotated data. Regarding the classification tasks, we found that training on such a small corpus produced very good results. Compared to other latent features difficult for automatic prediction such as narrativity or aspect markers in Chinese, the component of aspect that we examined seems more feasible to learn. We obtained results around 15% better than those for narrativity prediction (Meyer et al., 2013) for instance.

Grisot (2015) points out that according to a mixed statistical model estimated on the manually annotated corpus of 435 sentences, French tense is significantly determined by the interaction between narrativity and boundedness. Such theoretical insight is unfortunately hard to test empirically and will be investigated in further work. Using the same method presented in this paper, we can make two suggestions for using the information about the interaction between *narrativity* and *boundedness*. A classifier could be built to predict the narrativity and boundedness at the same time, i.e. a four classes task (+narrative+bounded, +narrative-unbounded, -narrative+bounded, -narrative+unbounded). The factored model would thereafter have one factor. Another solution would be to train two classifiers, one for narrativity and another for boundedness. This would produce two pairs of independent labels and hence two different factors in the factored model. It should be tested whether diluting the information in such a way would still add knowledge to the system, since the distributions may become scarce.

Acknowledgments

We are thankful to the Swiss National Science Foundation for partially founding the research presented in this paper through the COMTIS and MODERN projects (CRSI22-127510, 2010-2013, CRSII2-147653, 2013-2016). In addition, this research was completed while SL was supported by the Swiss National Science Foundation under grant no. P1GEP1_161877.
`http://www.idiap.ch/project/comtis`
`https://www.idiap.ch/project/modern`

(Koehn, 2010).

References

Bohnet, Bernd, Joakim Nivre, Igor Boguslavsky, Richárd Farkas, Filip Ginter, and Jan Hajič. 2013. Joint morphological and syntactic analysis for richly inflected languages. *Transactions of the Association for Computational Linguistics* 1:415–428.

Bojar, Ondřej, Christian Buck, Christian Federmann, Barry Haddow, Philipp Koehn, Johannes Leveling, Christof Monz, Pavel Pecina, Matt Post, Herve Saint-Amand, Radu Soricut, Lucia Specia, and Aleš Tamchyna. 2014. Findings of the 2014 workshop on statistical machine translation. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, WMT14, pages 12–58. Baltimore, Maryland: Association for Computational Linguistics.

Bojar, Ondřej, Rajen Chatterjee, Christian Federmann, Barry Haddow, Matthias Huck, Chris Hokamp, Philipp Koehn, Varvara Logacheva, Christof Monz, Matteo Negri, Matt Post, Carolina Scarton, Lucia Specia, and Marco Turchi. 2015. Findings of the 2015 workshop on statistical machine translation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, WMT15, pages 1–46. Lisbon, Portugal: Association for Computational Linguistics.

Carletta, Jean. 1996. Assessing agreement on classification tasks: The kappa statistic. *Computational Linguistics* 22(2):249–254.

Comrie, Bernard. 1976. *Aspect: An Introduction to the Study of Verbal Aspect and Related Problems*. Cambridge: Cambridge University Press.

Comrie, Bernard. 1985. *Tense*. Cambridge: Cambridge University Press.

Declerck, Renaat. 1991a. *Comprehensive Descriptive Grammar of English*. Tokio: Kaitakusha.

Declerck, Renaat. 1991b. *Tense in English: Its structure and use in discourse*. London: Routledge.

Declerck, Renaat. 2006. *The Grammar of the English Verb Phrase*, vol. 1 of *The Grammar of the English Tense System*. Berlin: Mouton de Gruyter.

Depraetere, Ilse. 1995a. The effect of temporal adverbials on (a) telicity and (un) boundednesss. In B. P.-M., V. Bianchi, O. Dahl, and M. Squartini, eds., *Temporal reference, aspect and actionality*. Turin: Rosenberg and Sellier.

Depraetere, Ilse. 1995b. On the necessity of distinguishing between (un)boundedness and (a)telicity. *Linguistics and Philosophy* 18(1):1–19.

Dowty, David. 1979. *Word Meaning and Montague Grammar*, chap. Word meaning and Montague grammar: The semantics of verbs and times in generative semantics and in Montague's PTQ. Springer.

Eisele, Andreas and Yu Chen. 2010. Multiun: A multilingual corpus from united nation documents. In *Proceedings of the 7th International Conference on Language Resources and Evaluation*, pages 2868–2872. European Language Resources Association (ELRA).

Gojun, Anita 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.

Gong, Zhengxian, Min Zhang, Chewlim Tan, and Guodong Zhou. 2012a. Classifier-based tense models for SMT. In *Proceedings of the 25th International Conference on Computational Linguistics*, COLING 2012, pages 411–420. Mumbai, India: The COLING 2012 Organizing Committee.

Gong, Zhengxian, Min Zhang, Chewlim Tan, and Guodong Zhou. 2012b. N-gram-based tense models for statistical machine translation. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, EMNLP-CoNLL'12, pages 276–285. Jeju Island, Korea: Association for Computational Linguistics.

Grisot, Cristina. 2015. *Temporal reference: empirical and theoretical perspectives. Converging evidence from English and Romance*. Ph.D. thesis, Université de Geneve, Geneva, Switzerland. Manuscript.

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

Hardmeier, Christian, Joakim Nivre, and Jörg Tiedemann. 2012. Document-wide decoding for phrase-based statistical machine translation. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 1179–1190. Jeju Island, Korea: Association for Computational Linguistics.

Hardmeier, Christian, Sara Stymne, Jörg Tiedemann, Aaron Smith, and Joakim Nivre. 2014. Anaphora models and reordering for phrase-based SMT. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 122–129. Baltimore, Maryland, USA: Association for Computational Linguistics.

Heafield, Kenneth. 2011. KenLM: faster and smaller language model queries. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 187–197. Edinburgh, Scotland, UK: Association for Computational Linguistics.

Koehn, Philipp. 2004. Statistical significance tests for machine translation evaluation. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing, A meeting of SIGDAT, a Special Interest Group of the ACL, held in conjunction with ACL 2004*, EMNLP 2004, pages 388–395. Barcelona, Spain.

Koehn, Philipp. 2010. *Machine Translation*. Cambridge: Cambridge University Press.

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

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

Loáiciga, Sharid, Thomas Meyer, and Andrei Popescu-Belis. 2014. English-French verb phrase alignment in Europarl for tense translation modeling. In *Proceedings of the 9th International Conference on Language Resources and Evaluation*, LREC'14, pages 674–681. Reykjavik, Iceland: European Language Resources Association (ELRA).

Manning, Christopher 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.

Meyer, Thomas. 2014. *Discourse-level Features for Statistical Machine Translation*. Ph.D. thesis, École Polytechnique Fédérale de Lausanne, Lausanne, Switzerland.

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

Och, Franz Josef. 2003. Minimum error rate training in statistical machine translation. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics - Volume 1*, ACL '03, pages 160–167. Stroudsburg, PA: Association for Computational Linguistics.

Olsen, Mari, David Traum, Carol Van Ess-Dykema, and Amy Weinberg. 2001. Implicit cues for explicit generation: Using telicity as a cue for tense structure in a chinese to english mt system. Tech. Rep. LAMP-TR-070,CS-TR-4248,UMIACS-TR-2001-33, University of Maryland, College Park.

Papineni, Kishore, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, ACL'02, pages 311–318. Philadelphia: Association for Computational Linguistics.

Reichenbach, Hans. 1947. *Elements of symbolic logic*. New York: Mcmillan.

Schmid, Helmut. 1994. Probabilistic part-of-speech tagging using decision trees. In *Proceedings of International Conference on New Methods in Language Processing*. Manchester, UK.

Song, Xingyi, Trevor Cohn, and Lucia Specia. 2013. BLEU deconstructed: Designing a better MT evaluation metric. *International Journal of Computational Linguistics and Applications* .

Vendler, Zeno. 1957. Verbs and times. *The Philosophical Review* 66(2):143–160.

Witten, Ian H., Eibe Frank, and Mark A. Hall. 2011. *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann.

Ye, Yang, Victoria Li Fossum, and Steven Abney. 2006. Latent features in automatic tense translation between chinese and english. In *Proceedings of the Fifth SIGHAN Workshop on Chinese Language Processing*, pages 48–55. Sydney, Australia: Association for Computational Linguistics.

Ye, Yang, Karl-Michael Schneider, and Steven Abney. 2007. Aspect marker generation for english-to-chinese machine translation. In *Proceedings of the Eleventh Machine Translation Summit*, MT SUMMIT XI, pages 521–527. Copenhagen, Danmark.

Zhang, Ying, Stephan Vogel, and Alex Waibel. 2004. Interpreting BLEU/NIST scores: How much improvement do we need to have a better system. In *Proceedings of the 4th International Conference on Language Resources and Evaluation*, LREC 2004. Lisbon, Portugal: European Language Resources Association (ELRA).