Chapter 7
Application to Natural Language Processing and Machine Translation

7.1 Temporal Cohesion Ties and Automatic Processing of Language

7.1.1 Natural Language Processing

Computational linguistics (CL), natural language processing (NLP) and machine translation (MT) are domains whose perspective on natural language is different from that of linguistic fields such as semantics, pragmatics and syntax. Their general purpose is to recreate automatically what humans naturally create—that is, produce and understand language. In neurolinguistics and psycholinguistics, there is a strong relation between linguistics and CL, NLP and MT. In particular, the automatic processing of language bases its models on linguistic theories qualitatively describing the functioning of human language, as well as on large quantities of data and the frequent behaviour of linguistic expressions. Language models developed in CL, NLP and MT find patterns of linguistic expressions and semantic interdependencies, allowing researchers to generalize behaviour, such as the parallel between temporal and pronominal reference suggested by Partee (1973) and further developed within the CL framework by Webber (1988). Well-known works on discourse structure and lexical aspect, such as Dowty (1979, 1986), Moens and Steedman (1987, 1988), Steedman (1997) and Moens (1987), among many others, were created within CL framework.

In the past few years, the literature on the processing of temporal reference has focused on issues such as event ordering (events relative to one another), time stamping (i.e. the temporal anchoring of a situation) and the generation of words expressing temporal relations for individual languages, usually for English. In this section, I will describe three principal existing studies in the NLP field related to temporal information. The first is a computational model of the semantics of Tense

C. Grisot, *Cohesion, Coherence and Temporal Reference from an Experimental Corpus Pragmatics Perspective*, Yearbook of Corpus Linguistics and Pragmatics, https://doi.org/10.1007/978-3-319-96752-3_7

and Aspect (Passonneau 1988). The second is a model for processing and automatically annotating temporal information in discourse, namely the TimeML annotation scheme model proposed for English by Pustejovsky and colleagues (2005a, b), and adopted by Bittar (2010) for French. The third is Li et al.'s (2001, 2004) model for processing Chinese.

Passonneau (1988) describes a processing system called PUNDIT, which processes references to situations and the intervals over which they hold using an algorithm that integrates the analysis of verbal tenses (i.e. Tense) and aspectual information (i.e. Aspect and Aktionsart). The algorithm was developed for English texts. Information from Tense and Aspect (perfective/perfect[1] or progressive) as well as temporal adverbials such as *before, after* and *when* is used to derive three complementary pieces of information:

- Determine whether the situation is associated with the realis or irrealis world. Different processes are needed, according to whether the situation refers to actual or potential time.
- Determine the internal temporal structure of the predicated situation—i.e. the inherent temporal information of the verb phrase—as one of three situation types: *state, process* and *transition event* (the third referred to as achievement and accomplishment in Vendler's terminology).
- Determine the temporal localization of the actual situation with respect to the moment of speech/text production, or to the times of other situations, with the help of Reichenbachian temporal coordinates E, R and S.

The internal temporal structure of a situation consists of one or more intervals. Each interval is characterized by two features, *kinesis* and *boundedness*. Kinesis pertains to the internal structure of an interval, and can be *stative* or *active*. Stative kinesis signifies that "each subinterval is equivalent to any other subinterval with respect to the asserted situation" (Passonneau 1988, 47). Processes and transition events have active kinesis involving change from one subinterval to another. Boundedness relates to whether or not an interval is bounded, and constrains the manner in which the situations are located in time (i.e. temporal reference). The intervals associated with states are inherently unbounded, although they can become bounded by an appropriate temporal adverbial. Processes (activities in Vendler's terms) are generally unbounded, and can become unspecified for boundedness if the verb is progressive. In (618), the time shown by the clock is interpreted as falling within the unbounded interval of *ringing*, but in (619), where the verb is not progressive, it can be interpreted as marking the inception of the process or roughly locating the entire process (Passonneau 1988, 47).

(618) The alarm was ringing at 8 am.
(619) The alarm rang at 8 am.

[1] The model uses the term *perfect* to refer to the English Present Perfect and Past Perfect verbal forms. Perfect verbal forms (*relative* in Reichenbach's terms) have an R distinct from E.

Table 7.1 Module 3: temporal localization

Parameter	Value	Rules
Perfect	Yes	E < R
	No	E = S
Tense	Past	R < S
	Present	R = S

These temporal pieces of information are assembled in a context-dependent compositional semantics framework. Passonneau points out the complexity of computing temporal information from several sources, since the contribution of each distinct component can depend upon co-occurring elements. Her suggestion is a model of extracting temporal information by separating temporal analysis into distinct tasks, each task targeting one type of temporal input. Each task provides input for the next stage of analysis, and this must be provided as explicitly as possible to avoid conflicting with the subsequent processing. The algorithm for the temporal analysis of an inflected verb contains three modules. The first module computes the actual time (realis) from temporal information provided by Aspect, Aktionsart and Tense. Only realis sentences are considered for further analysis. The second module derives the inherent temporal structure of the situation from two temporal parameters, lexical aspect and progressive aspect. The output of the second module—that is, an explicit representation of the situation's temporal structure and the event time—is sent to the third module, which derives the temporal localization of the situation from the last two parameters, perfect verbal form and tense. Temporal localization is established with the help of Reichenbachian temporal coordinates. However, the model diverges from Reichenbach, primarily by distinguishing between the event time and the temporal structure of a situation (Passonneau 1988). Module three is illustrated in Table 7.1.

The possible combinations of the values of all the parameters considered are provided in Table 7.2. A situation is thus located in time in reference to the parameters of Aspect, Aktionsart and Tense, and its interpretation depends on this temporal localization. The Simple Present locates unbounded temporal structure coinciding with S, while processes and transition events do not refer to the actual moment of speech of the utterance, as shown by the interpretation of (620). The Simple Past locates the event time of any temporal structure prior to S. However, each temporal structure provides differences in interpretation regarding the surroundings of the event time. Perfect verbal forms provide more supplementary information than simple forms, specifically about the relation between R and E.

(620) The pump operates.

To the best of my knowledge, Passonneau's account of temporal information in discourse is the first model to integrate semantic information from several linguistic sources. Another semantic account of temporal information, called the Specification

Table 7.2 Possible combinations of temporal localization of situations

Tense	Aspect	Stative	Process	Transition event	Location
Present	Simple	Unbounded	Not actual time	Not actual time	E = S = R
	Perfect	Unbounded	Unspecified	Unspecified	E < R = S
	Progressive	Unbounded	Unbounded	Unbounded	E = S = R
Past	Simple	Unbounded	Unspecified	Bounded	E = R < S
	Perfect	Unbounded	Unspecified	Bounded	E < R < S
	Progressive	Bounded	Unbounded	Unbounded	E < R = S

Language TimeML, was developed in the AQUAINT[2] programme. TimeML is a semantic annotation framework for temporal information in discourse, and provides guidelines for trained humans who carry out the annotation (Pustejovsky et al. 2005a, b).[3] TimeML was designed to address four issues regarding temporal information:

- Temporal localization of situations (identification and anchoring in time)
- Ordering of situations with respect to one another (lexical and discourse ordering)
- Dealing with contextually underspecified temporal expressions (such as *last week* or *2 weeks before*)
- Dealing with the persistence in time of situations

TimeML considers all temporal objects in a discourse, broadly grouped into *temporal expressions* (adverbials and connectives) and *events*. The class of *events,* which includes *inflected verbs* and *event nominals*, is a generic term used for verbs describing various types of states and events. It makes reference to Reichenbach's (1947) description of verbal tenses, Vendler's (1957, 1967) aspectual classes, the distinction between *lexical* and *grammatical* aspect, and Bach's (1986) notion of *eventualities*. The annotation language consists of a set of *basic tags* for expressing events, explicit temporal expressions and function words, and a set of *links* between the annotated elements, which have different types, such as *temporal, subordination* and *aspectual*.

The <EVENT> tag is used to annotate both inflected verbs (predicative and non-predicative tenses) and events expressed by nouns. Verbal tenses are expressed in terms of a combination of Tense (with a choice between *present*, *past* and *future*) and Aspect (with a choice between *progressive, perfective, progressive-perfective* and *none*). Verbs are categorized into seven classes: *reporting, perception, aspectual,*

[2] The AQUAINT programme is a dedicated effort to improve the performance of question answering systems using free text available on the Web. An important aspect of this research is its access to information from text by way of content rather than keywords. It aims to create a specification language to identify events and temporal expressions in text.

[3] The TimeML framework adopts XML as formal language, and provides a formalized markup language called ISO-TimeML, with a systematic means of extracting and representing temporal information. The annotation framework's specification and guidelines are available at http://timeml.org/site/publications/specs.html

states, demanding an action, demanding a state and *occurrences.* These classes are relevant due to the type of relation (link) they require. The tag <TIMEX3> is used to mark up explicit temporal expressions referring to *day times, dates, durations* and *sets.* The tag <SIGNAL> is used to annotate function words, which indicate how temporal objects are to be related to one another. Signals are generally: temporal prepositions (*on, in, at, from, to, before, after, during,* etc.); temporal conjunctions (*before, after, while, when*, etc.); and special characters ("-" and "/" in temporal expressions denoting ranges, such as *September 4–6, April 1999/July 1999*, etc.).

The tags <TLINK>, <SLINK> and <ALINK> serve to capture the different types of relations existing between two events (in the broad sense used in this framework), and between an event and an explicit temporal expression. These links can have a *temporal* nature (such as *before, after, includes, simultaneous, during, identity,* etc.), a *subordination* nature (such as *evidential, factive, counter-factive, conditional,* etc.) and an *aspectual* nature (such as *initiates, culminates, terminates, continues,* etc.).

Example (621) shows a sentence and its interpretation in TimeML, paraphrased in the following terms: the temporal adverb *today* is annotated with the tag TIMEX3, which expresses a date and has the identification tag t32; there is a temporal link with the value *before* between the event number 2 from the sentence and this adverbial, shown by the TLINK tag at the end of the formal description. Two events are mentioned in the sentence: the first is expressed by the verb *learned* (which is described as a reporting verb, expressing past tense); the second is expressed by the verb *has taken* (which is described as an occurrence verb, expressing present tense and the perfective aspect). This kind of annotation, carried out by trained humans, allows the explicitation of temporal information that is implicit at the discourse level.

(621) Finally, today we learned that the space agency has finally taken a gian leap forward.
Finally today,*we learned that the space agency has finally taken a giant leap forward.*

The metadata markup language TimeML is therefore a formal framework which integrates three types of semantic temporal information: (i) the temporal anchoring of situations with respect to S and R; (ii) the temporal ordering of situations relative to one another, both intrasententially and in discourse; and (iii) the semantics of underspecified temporal expressions, by integrating them in the overall interpretation of the discourse. Corpora manually annotated with the TimeML language are useful tools for finer-grained analyses of temporal information. TimeML is an important example of the efforts made by researchers to integrate temporal information from several sources, and to make explicit the various types of relations existing between situations. However, as I have argued in Sect. 2.3, temporal information cannot be processed according to linguistic or semantic sources alone.

Both Passonneau's model and TimeML are models developed for tensed languages, such as English and French. Li et al. (2001, 2004) developed a model for processing temporal reference in Chinese. They report on a computational model

based on machine learning algorithms. The core model consists of a set of rules combined with a set of linguistic features for the purpose of temporal relation resolution. The linguistic features used are Chinese words which can function as temporal indicators—time words (e.g. *year, month)*, time position words (e.g. *a few days ago*), temporal adverbs (e.g. *lately, recently),* auxiliary words and verbs, aspectual markers (e.g. *le, zhe* and *guo*), prepositions and special verbs, among others. Temporal relations are described in terms of E, R and S (Reichenbachian coordinates). The TICS system (Temporal Information-extraction from Chinese Sources) receives financial texts as input, analyses each sentence one by one in order to extract temporal information, and represents each piece of information in a concept frame. All concept frames are linked according to the temporal relations holding between events. This model points out NLP models' need to identify temporal relations holding between eventualities, in order to have accurate results.

To sum up, in this section I have discussed three NLP studies on the automatic processing of temporal information at the discursive level, and shown that automatic systems make use of temporal information from various linguistic sources: verbal tenses, grammatical and lexical aspect, the location of eventualities with respect to Reichenbachian coordinates E, R and S, temporal adverbials, and other linguistic markers, especially relevant in tenseless languages.

7.1.2 Machine Translation

In the MT field, two main types of automatic translation systems exist: *rule-based* and *statistical* systems. Rule-based systems were the first to be created in the 1970s, such as the 'pioneer' Systran company (currently a hybrid between rule-based and statistical system), the Logo company, and the MT system developed at the University of Montréal for weather forecast translation. In the 1980s, important research was carried out on the English/Japanese language pair, while the subsequent German Verbmobil project in the 1990s had some success in speech-to-speech translation (for a more detailed discussion, see Meyer 2014, Chapter 1). For these systems, a large set of lexical and/or syntactical rules had to be written by linguists and manually implemented. As pointed out by Meyer (2014), this costly procedure made it hard to adapt these systems to other language pairs, directions of translation, or stylistic registers. The functioning of rule-based systems is designed to have three levels. The first and bottom level consists of translation word-by-word, with the possible re-ordering of words. At the second and middle level, the system operates at the syntactic level via transfer rules, implemented on syntax trees, from a source language to a target language. The third and most complex level is creating by building an *interlingua*, which is a 'completely language-independent semantic representation of the source text's meaning' used to generate the target text directly (Meyer 2014, 4). However, building the interlingua proved to be a very problematic task because of the difficulty of integrating world and domain knowledge.

As a result, most of the research on MT throughout the 1990s focused on statistical systems. In SMT, where there is no rule-based processing, the goal of the system is to learn the correct translations of words, phrases and sentences from large corpora translated by humans—i.e. parallel corpora that nowadays exist in several languages, such as EuroParl (Koehn 2005). The most common SMT system is the *phrase-based* system,[4] which is the product of several components, none of which involves linguistic knowledge. The first component is a *phrase translation model,* trained on aligned (both at sentence- and word-level) parallel corpora, which computes translation probabilities for all sequences of words in the source text. The second is a *language model*, which specifies the probability of the string of words considered by the SMT system, as well as syntactic and lexical information of the target language—in other words, estimating how much a candidate translation conforms to fluent target language. The third is 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, creating several hypothesis translations, and ultimately chooses the best one according to the language model and the reordering model (Koehn 2010).

The functioning of an SMT system can be described in three stages. The first is the *training* stage, in which the system learns the most likely correspondences, reordering the chunks of words from parallel corpora. The second is the *tuning* stage, in which the system is trained on a much smaller text, ideally of the same register as the target text, in order to optimize the language pairs identified in the first stage. The third is the *testing* stage, in which a new text is handed to the system for translation. In this stage, the system tries to find the most likely phrase pairs, and recombines these hypotheses based on probability scores from the translation and the language model available. One of the most often used, freely accessible statistical MT systems is Google Translate.

Other attempts to improve the results of SMT systems were mainly made to include linguistic information in the system itself. Two of them were to create hybrid systems using both linguistic rules and statistical methods (such as Systran, Reverso and Linguatec), and to use additional knowledge within the SMT paradigm. For the latter, researchers proposed *factored translation models* (Koehn and Hoang 2007), which are usually used to add morphological, semantic or pragmatic information. This information is provided to the system via annotation of the parallel data. The training data is enriched with the desired linguistic information, and is automatically annotated by a *classifier*. A classifier is a tool that makes use of

[4] SMT systems use word or phrase alignment algorithms to align the words of a sentence in two languages, the source and the target. There are four types of alignment (Samardzic 2013, 94–95): (i) *one-to-one* (when corresponding single words are identified, i.e. pairs of words); (ii) *one-to-null* (used to describe words that occur in one language where no correspondent can be found in the other language); (iii) *one-to-many* (when one word in a language corresponds to several words in the other language); and (iv) *many-to-many* (when no single word is an alignment unit). The first three types are called word-based alignments; the last is called phrase-based alignment.

machine learning algorithms,[5] usually according to human-annotated data, taking data items and placing them in one of the available classes. One type of classifier is the *maximum entropy* (MaxEnt) classifier, which can be built with the Stanford Classifier Package (Manning and Klein 2003). The underlying principle of maximum entropy is that, when assigning a class where there is no external knowledge, one should prefer uniform distributions, and thus assign the considered classes uniformly. Annotated data used to train these classifiers provide external knowledge, thereby informing the automatic labelling technique where to be minimally non-uniform (i.e. where not to provide uniform distributions of the tags). Iterative application of the classifier result in automatically labelled or annotated texts with the features considered. The classifier plays a crucial role in an SMT system, because it automatically produces tags that increase the probability that a certain string of words in a target language is the correct translation. For this reason, much work has been done on the construction of the classifiers, such as the research carried out in the COMTIS and MODERN Swiss research projects (cf. Introduction, footnote 1), which focused on Western European tense-prominent languages. I will discuss this research in Sect. 7.2.

A series of studies in MT have focused on the automatic translation of temporal information in general, and of verbal tenses in particular. Most of them (such as Olsen et al. 2001, Ye et al. 2006, 2007) are on the Chinese/English pair of languages, due to the typological differences betweent the two languages (tenseless for the former and tensed for the latter). Olsen et al. (2001) and Ye et al. (2006) aimed to improve machine translation from Chinese to English; Ye et al. (2007) were interested in machine translation from English to Chinese. The different strategies used to encode temporal information in English and Chinese are challenging for the automatic translation of tense and aspect. Ye and colleagues point out that neither word-based alignment nor phrase-based alignment can capture the mapping between the tense markers in English (morphemes) and the aspect markers of the corresponding Chinese verbs (lexemes). For example, when Chinese aspect is marked, it takes the form of a separate word, such as the *le* marker, which aligns poorly with English tensed verbs, and so the aspectual information is dropped. As a result, instead of producing (622), SMT systems produce the sentence in (623), using the infinitive form of the verb and, in this case, with a different lexical choice (Loáiciga and Grisot 2016, 8).

(622) Wo ji le yi feng xin gei ta
I send PERF one QUANTIFIER letter to/for he
'I sent him a letter.'

(623) Wo xie yi feng xin gei ta.
I write one QUANTIFIER letter to/for he
'I write him a letter.'

[5] Samardzic (2013, 112) explains that the data which machine learning algorithms take as input are considered as *experience*. A computer programme "learns from experience" if its performance with respect to a task improves with experience—i.e. by dealing with the data.

Olsen et al. (2001) used information about Aktionsart—in particular, the *telicity* ontological feature—in order to predict whether the verbal tense expresses reference to present or past time in the target language, English. They built a system (interlingua model) which allows them to obtain reliable lexical information associated with each verb. Their hypothesis is that Chinese sentences with a telic aspect will translate into English as 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. The results are given in terms of accuracy, or correct translations. While the baseline system (unware of telicity) reached 57% correct translations, a second system which uses the telicity property of verbs reached 76% correct translations. Furthermore, a third system, built using telic information alongside other linguistic information such as Aspect and adverbials, reached 92% accuracy. Their system is highly deterministic, with a fixed correspondence *+telic* → reference to *past*, *−telic*→ reference to *present*. However, this deterministic correspondence might not applicable to other pairs of languages, and the identification process of telic verbs relies heavily on their particular system's lexicon, making it difficult to implement in different systems.

Ye et al. (2006) built a classifier that generates tense marking in English. The classifier learns the mapping between English and Chinese from a set of features provided by a training set of data. Since the purpose of the SMT system is to translate into English, they used features of English to predict tense marking. Their main argument is that NLP work must aim to build systems that follow the mechanisms of the human brain, in order to optimize their performance. In their words (2006, 50):

> The bottleneck in Artificial Intelligence is the unbalanced knowledge sources shared by human beings and a computer system. Only a subset of the knowledge sources used by human beings can be formalized, extracted and fed into a computer system.

The features based on knowledge shared with human beings are called *latent features*. Olsen et al. (2001) illustrated the value of latent features by showing how lexical aspect or the telicity of the verb phrase improve the translation of temporal reference from Chinese to English. Ye et al. (2006) used several surface features (formal features) and two latent knowledge sources, namely *telicity* as proposed by Olsen et al. (2001), and *event ordering* as implemented in the TimeML annotation scheme. The surface features used to generate tense markers in English are (2006, 50):

- The type of speech act.
- The syntactic structure in which the current verb is embedded.
- The occurrence of temporal adverbials and aspectual markers.
- The distance in number of characters between the current and the previous verb, and whether the two verbs are in the same clause or not.

The two latent features are assumed to be used by human beings in tense resolution (though psycholinguistic and neurolinguistic studies have only recently begun to investigate them). Information about the lexical aspect is used in terms of *telicity*

(i.e. the verb's ability to be bound within a certain time span) and *punctuality* (i.e. punctual verbs, or achievements in Vendler's terms). The authors point out that a verb's telicity value is context-dependent. The second latent feature concerns the temporal relations holding between eventualities. The authors defined temporal relations in terms of precedence, inclusion, overlapping and lack of temporal relation. As such, they used human-annotated data with these two latent features. The classifier trained on surface and latent features had significantly better results (83.4% accuracy) than the classifier trained only on surface features (75.8%) and the classifier trained only on latent features (80%).

Ye et al. (2006) provided evidence that lexical aspect and the temporal relations holding between eventualities are significant factors in predicting verbal tenses in a target language. In this research, specifically in Sect. 4.4, I suggest a model that uses several latent features, such as Aspect, Aktionsart, and temporal and causal relations holding between eventualities (grouped under the [±narrativity] feature encoded by Tense) to predict the verbal tense in several target languages. The advantage of the research presented in this book, compared to previous models for SMT, is that all features are captured automatically.

Ye et al. (2007) report the building of a classifier that generates aspectual markers in Chinese: *le, zhe*, *guo,* and NULL when none of the three occurs. Since the purpose of the SMT system is to translate into Chinese, the features used to predict aspect marking correspond to both English and Chinese. Five surface features and one latent feature (2007) were used:

- Syntactic features, which can influence the verb's tendency to take an aspectual marker.
- Positional features, indicating that the occurrence of a verb with another can influence the verb's tendency to take an aspectual marker.
- Signal lexeme features, indicating that the aspectual markers considered present certain lexical occurrence patterns (for example, with some auxiliary words and not with others).
- A phonological feature, indicating that aspectual markers are incompatible with idioms that have four Chinese characters.
- An English verbal tense feature, indicating that verbal tenses play the same role as aspectual information in Chinese, i.e. expressing temporal reference.
- Lexical aspectual features, pointing to the theoretical assumption that the inherent features of the verb phrase play an important role in establishing temporal reference.

Verbal tense in English and lexical aspect have been manually annotated. The classifier's performance was significantly better than a simple classifier, which always assigns the most frequent aspect marker (which is *le*). All features used to predict aspectual markers in Chinese were significant, but behaved differently for each of the three aspectual markers considered. For example, the lexical aspectual features was only significant for the prediction of the aspectual marker *zhe*, whereas the English verbal tense feature was significant for predicting the occurrence of *le* and NULL. These two studies involving the translation from and into Chinese, a

tenseless language, point to the fact that dividing temporal information from Tense, Aspect and Aktionsart and using it as latent features is useful for improving the translation of a text with respect to temporal reference.

Work on the English-French pair of languages has been done by Loáiciga et al. (2014), as well as Meyer et al. (2013), and Loáiciga and Grisot (2016). Loáiciga et al. (2014) automatically identified all English and French verb phrases in EuroParl, which is a large parallel and aligned corpus. They then automatically annotated the verb phrases on both sides of the corpus with one of 12 verbal tenses, indicating reference to present, future or past time. The annotation allowed them to map and to measure the distribution of tense translation between the languages. They found that the ambiguity of the translation of the English Simple Past into the French Passé Composé, Imparfait, Passé Simple and Présent is statistically significant ($p < 0.05$).

Using this automatically annotated corpus, the authors present two SMT experiments on disambiguating the translation of the English Simple Past into French. Firstly, the parallel and aligned corpus is used to annotate the English verb with the French tense automatically. For example, 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 trained an SMT system on this annotated corpus, securing an increase of 0.50 BLEU[6] points over a baseline with no French verb tense labels.

In a second experiment, the authors used the corpus to train a classifier of French verb tenses using features from the English component only. In other words, the information regarding the French verb tense is not used, and tense labels are instead predicted. In Loáiciga and Grisot (2016), we point to the fact that this classification task is not trivial, since it involves nine classes corresponding to nine verbal tenses (all four *future* and *conditional* tenses of the original 13 tenses were grouped together into one single class) inferred from the source language. Results vary significantly depending on the particular verbal tense, ranging from an F1 score[7] of 0.16 for the Passé Simple to 0.46 for the Imparfait, 0.77 for the Passé Composé and 0.92 for the Présent. Finally, they provide the SMT system with the French tense labels produced by the classifier, and therefore prone to error. This second system

[6] The BLEU score (Bilingual Evaluation Understudy; Papineni et al. 2002) counts the overlap in terms of matching number of words and n-grams between the candidate translation and one or more reference translations. The more matches there are for 4-, 3-, 2- and 1-grams in a candidate translation compared to its reference, the higher the BLEU score. The values of the score range from 0 to 100, where the latter indicates identical translations. The existing SMT systems usually have scores between 11-33 BLEU points. BLEU is accepted as the best metric in terms of matching human judgments of translation quality, especially when averaged over a large quantity of text.

[7] The metrics used in computational linguistics to evaluate classification results are: *accuracy* (percentage of correctly classified instances); *precision* (percentage of correctly classified instances among correctly identified instances); and *recall* scores (percentage of correctly classified instances over all instances) (Meyer 2014, 50). Precision and recall correspond to Type I and Type II errors in statistics, and are used (their harmonic mean) to determine the F1 score, which ranges from 0 (worst score) to 1 (best score).

performance increased by 0.12 BLEU points over the baseline. They note that the quality of the translation was determined to a great extent by the quality of the classifier for each particular verbal tense. For example, for translating the Imparfait and the Subjonctif, the second system (tense-aware) was much better than the baseline, whose results did not exceed statistical predictions based on the parallel corpora.

7.2 The Automatic Classification of [±narrativity] and [±boundedness]

One of the purposes of this research was to improve the results of a statistical machine translation (SMT) system when it comes to the translation of verbal tenses. Current SMT systems have difficulties in choosing the correct verb tense translations, because these depend on a wider-range context than SMT systems consider. SMT systems aiming to model intersentential relations, such as the temporal information conveyed by Tense, Aspect and Aktionsart, require large numbers of annotated corpora, with semantic and pragmatic information to be used in the training phase of the statistical system.

Large amounts of annotated data can either be produced manually or automatically. Unfortunately, manual annotation of large amounts of data is time consuming, and very expensive. For these reasons, manual annotation is usually performed on smaller amounts of data. As for automatic annotation, one can choose to use existing automatic tools dealing with temporal information in the discourse, such as the TimeML markup language, or to build a *classifier*. A classifier is trained on a small amount of annotated data, and learns the annotation scheme by way of machine-learning algorithms. The classifier is used thereafter to annotate large amounts of data, necessary for the SMT system.

At this point of the discussion on the type of data, one issue that is worth mentioning regards the trade-off between using a small quantity of accurate data (generally human-annotated or human-post-edited) on the one hand, and using a large quantity of imperfect data on the other hand. Large quantities of imperfect data can be used in so-called *on-line* and *unsupervised learning* (i.e. the system learns all the patterns emerging from the data), and are very useful in binary classifications for unambiguous cases. However, for ambiguous (and also underspecified) cases, which are difficult to classify, the usefulness of large quantities of imperfect data is limited. In such cases, human intervention is generally required in order to reach an accurate judgement. As such, small accurate quantities of data are necessary, especially for the classification of difficult cases, and are used in so-called *supervised* learning. This is the case for the annotation experiments with [±narrativity] and [±boundedness] features reported in this chapter. The choice of one of the two types of data depends on the task, and the two methods can be used to complement one another.

In the COMTIS and MODERN projects, two classifiers were built in order to annotate data automatically with labels learnt from human-annotated texts. The first classifier automatically annotates texts with the [±narrativity] feature (Grisot and Meyer 2014; Meyer 2014). The second classifier deals with the [±boundedness] feature (Loáiciga and Grisot 2016). Human annotation experiments with these two features were described in Sects. 4.2.7 and 4.3.2. I will describe the automatic annotation experiments in Sect. 7.2. Several SMT systems were built, trained on the data annotated by the two classifiers. The results of the MT experiments provided in Sect. 7.2.2 show that SMT systems which are aware of the linguistic information provided by annotation experiments (i.e. information about the temporal ordering of eventualities and about lexical aspect) translate verbal tenses more accurately, and make better lexical choices (Meyer et al. 2013; Loáiciga and Grisot 2016).

7.2.1 Automatic Annotation Experiments

In Sect. 7.1.2, I spoke about SMT systems targeting the English-Chinese pair of languages (Ye et al. 2006; Ye et al. 2007). In these studies, the classification results were not embedded in an SMT system, and the classifier classes were the actual verbal tenses. In Grisot and Meyer (2014) and Meyer et al. (2013), we use classification as a means of enhancing an SMT system with knowledge about the [±narrativity] feature in order to produce better choices of verbal tense when translating from English into French. In Loáiciga and Grisot (2016), we use knowledge about a pragmatic component of lexical aspect, the [±boundedness] feature, in order to produce better choices of verbal tense in French. These two features are two essential features from the HD model of temporal reference (Chap. 5).

The data used in the automatic annotation experiments consist of 435 English Simple Past items, initially used in the annotation experiments described in Sect. 4.2.7 with the [±narrativity] feature, and in Sect. 4.3.2 with the [±boundedness] feature. A classifier was built for each of these features, and trained on the human-annotated data. For each classifier, a series of surface features was considered.

7.2.1.1 Annotation of the [±narrativity] Feature

The training data contained 257 narrative and 178 non-narrative English Simple Past items (a total of 435). The performance of the classifier was tested on a smaller sub-portion of the corpus which had previously been annotated manually, with the same stylistic genre distribution, consisting of 118 items of the English Simple Past: 75 instances of narrative, and 43 of non-narrative. Surface features were obtained from syntactic and part of speech (POS) parsing of the verbs occurring in the experimental items, using Charniak and Johnson's constituent parser (2005), and temporal analysis of the text with the TimeML parser (Verhagen and Pustejovsky 2008). The surface features used were the following:

- Neighbouring verb word forms.
- The position of the verbal tense in the sentence.
- The POS tags of all the words in the sentence.
- The syntactic tree structure of the sentence.
- Temporal markers (such as *while, since, weeks/days after or before, subsequently, repeatedly* and the like) from a hand-made list of 66 temporal discourse markers, inspired by the temporal connectives annotated in the Penn Discourse Treebank (Prasad et al. 2004, 2008)
- The types of temporal marker (from TimeML), such as temporal *simultaneity* or *sequencing* for temporal markers, *infinite, participle* or *future* for the class of verbal tense, and *perfective* or *imperfective* for the grammatical aspect.

With these features, a MaxEnt classifier, built with the Stanford Classifier package (Manning and Klein 2003), achieves an F1 score of 0.72 (the weighted mean of precision and recall). Out of the 118 test instances, the classifier correctly annotates 90 items, corresponding to 76.3%. Moreover, the κ value for the agreement between the classifier and the reference is 0.46. The classifier was then used to label automatically the Simple Past verbal tenses in the English component of a large parallel corpus necessary to train an SMT.

In order to test the classifier's performance further, Meyer (2014, 76) reports that the disagreements occurring in the manual annotation experiment (cf. Experiment 3) were resolved by directly inferring the narrative/non-narrative labels from the verbal tense occurring in the French component of the parallel corpus in a deterministic manner: a Passé Simple or Passé Composé correlates with a narrative label, and an Imparfait with a non-narrative label. When trained on such data, the classifier only achieves an F1 score of 0.71, and has a κ of 0.43 in the test set, even though it was trained on more data than before. This confirms two points: the first is the score range that can be expected when trying to classify narrativity automatically, which is 0.46; the second is that narrativity cannot be correlated with French verbal tenses in a deterministic one-to-one correspondence.

In addition, Meyer (2014, 76) reports the construction of another classifier—the CRF model (Lafferty et al. 2001)—which labels narrativity in sequence with other tags, such as part-of-speech (POS) tags. The CRF uses the two preceding POS tags as features to label the next POS tag in a sequence of words. The same training set of 435 sentences as used above was POS-tagged using the Stanford POS tagger (Toutanova et al. 2003), and the tags of VBD given for Simple Past verbal tenses were replaced with narrativity labels from manual annotation. The same procedure was applied for the 118 sentences used for testing the performance of the classifier. The CRF classifier had a lower performance than the MaxEnt classifier: its correct labelling of narrativity only reached an F1 score of 0.36, with a negative κ value signalling a weak inverse correlation.

According to Meyer (2014), the performance of the MaxEnt classifier was boosted by the temporal and semantic features used as surface features, such as the manually created list of temporal connectives, and the type of temporal markers taken from TimeML (temporal *simultaneity* or *sequencing* for temporal markers,

infinite, participle or *future* for the class of verbal tense, and *perfective* or *imperfective* for the grammatical aspect). We can therefore conclude that this information is useful for enhancing narrative and non-narrative usages of verbal tenses, because they relate to the ConText in which the [±narrativity] procedural feature must be determined.

7.2.1.2 Annotation of the [±boundedness] Feature

As before, the Stanford Maximum Entropy package (Manning and Klein 2003) was used to build a MaxEnt classifier. The training data contained 435 Simple Past occurrences, judged by human annotators as bounded or unbounded in the experiment from Sect. 4.3.2. As such, the training data contained 236 bounded and 199 unbounded instances. In this experiment, the *cross-validation* method was used to determine the training and the testing data. This method consists in automatically splitting the data into several equal sub-parts (ten in this case, therefore a tenfold cross-validation). The classifier is trained iteratively on nine parts, and its performance is tested on the tenth part. Finally, the classifier's performance is calculated as the average of the results it had for each of the ten iterations.

This experiment used several additional features resulting from the annotation experiments described in Sects. 4.2.7 and 4.3.3, and from human editing of the data. Since this is a fully supervised classification partially fed with features known to be pertinent for the task, its results are expected to be a measure of the maximum success rate for this particular task. Two classes of features were used to enhance the classifier: syntactic and temporal features. Manually annotated features resulted from the human annotation experiments (the 435 Simple Past occurrences annotated in Sects. 4.2.7 and 4.3.3), indicated by a * symbol. For the automatically generated features, the dependency parser of Bohnet et al. (2013) from MateTools was used on the English component of the corpus to produce POS tags and dependency labels.

The syntactic features are as follows (Loáiciga and Grisot 2016):

- Simple Past token*: Simple Past instances to be classified, identified manually.
- Infinitive form*: the non-finite form of the English Simple Past.
- Grammatical aspect*: a binary feature, originating from the translation of the corpus into Serbian and its recovery by the cross-linguistic transfer of properties method, with two values (perfective and imperfective, cf. Sect. 4.3.3).
- French verbal tense*: identified in the French part of the translation corpus by the translation spotting technique.
- Position in the sentence: refers to the ordinal position of the English Simple Past verb in the sentence.
- POS-tags of the English Simple Past token: these distinguish between active voice Simple Past verbs, such as *went* (VBD), compound active voice Simple Past verbs such as *did go* (VBD + VB), and passive voice Simple Past verbs, such as *was taken* (VBD + VBN).

- The head and its type: this refers to the syntactic head of the verb to classify, along with its POS-tag.
- Children dependencies: these indicate the dependency relation of the three nearest children of the English SP verb.
- Children POS-tags: these 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, such as the presence of *in* prepositional phrases (e.g. *in 2 months*) for bounded eventualities.

The temporal features are as follows:

- Temporal markers (such as *while, since, weeks/days after or before, subsequently, repeatedly* and the like) from a hand-made list of 66 temporal discourse markers, inspired by the temporal connectives annotated in the Penn Discourse Treebank (Prasad et al. 2004, 2008)
- The types of temporal marker (from TimeML), such as temporal *simultaneity* or *sequencing* for temporal markers, *infinite, participle* or *future* for the class of verbal tense, and *perfective* or *imperfective* for the grammatical aspect).
- The [±narrativity] feature*: issued from human annotation experiments.

With these features, the classifier hits an F1 score of 0.89 for the bounded class, and 0.87 for the unbounded class, and has 88% accuracy. These scores indicate the classifier's very good performance. These results are partially explained by the features taken from the human annotations, signalled by the * symbol. The most informative features, in descending order, are: grammatical aspect; verbal tense used in French; narrativity; and the infinitive form of the verb in English. Of these features, grammatical aspect and narrativity (as well as boundedness with respect to its interaction with narrativity) also turned out to be significant in the mixed model adjusted to predict the verbal tense used in the target language (cf. Sect. 4.4).

In Loáiciga and Grisot (2016), we point out that, 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. The gold (human) annotation of the bounded and unbounded labels was not perfect. The κ value for the inter-annotator agreement rate in Experiment 4 was 0.84, which is already much higher than in Experiment 3, on the [±narrativity] feature.

In Loáiciga and Grisot (2016), we present a second experiment, in which certain surface features were generated automatically from raw data, such as the Simple Past token, the infinitive form, the position in the sentence, the POS tags of the verbs, and the POS tags of their arguments (the verb phrase). Three features originating from the human-annotated data were not used in this experiment: the [±narrativity] feature; grammatical aspect; and French verbal tense. Since human-annotated data is costly and time-consuming, this second experiment aimed to test whether the classifier has reliable results if it is trained only on automatically extracted surface features, which might have errors. Consequently, the results of this experiment are expected to give a realistic impression of the quality of detecting boundedness in a

large corpus using automatically generated features and a small quantity of annotated data (the only annotation being the gold prediction class) for training.

In the previous experiment, a MaxEnt classifier was built. The dependency parser of Bohnet et al. (2013) and the TreeTagger (Schmid 1994) producing POS-tags and lemmas were used on the English component of the corpus. With automatically-generated features, the classifier hits an F1 score of 0.84 for the bounded class and 0.79 for the unbounded class, with 82% accuracy. The results represent the average classification using ten-fold cross-validation. Compared to the first experiment, these scores still represent reliable results, and they show that both the bounded and the unbounded category are more difficult to predict solely according to automatically generated features. The difference of approximately 8% for each category between the results of the two classifiers is shown to be statistically significant by a two-sided t-test (τ (434) = 7.28, $p < .05$) This result can be interpreted in terms of the quality of human-annotated data compared to automatically generated data, which contains a percentage of errors. Nevertheless, the second classifier was still able to learn how to discriminate between bounded and unbounded Simple Past occurrences in a satisfactory manner.

For a more precise image of the importance of using linguistic information for SMT systems, Loáiciga and Grisot (2016) set a baseline based on the random distribution of bounded and unbounded labels in the corpus, 54% for the former and 46% for the latter (cf. the experiment from Sect. 4.3.2). A random sample with resampling of 435 bounded/unbounded labels was generated, with probabilities of 0.54 and 0.46 respectively. The random labels obtained were compared to the human-annotated corpus, in order to compute precision, recall and F-score. The random sample has an F1 score of 0.56 for the bounded class and 0.47 for the unbounded class, and has 54% accuracy. The results of both the classifier using human-annotated features (experiment 1) and the classifier using only automatically-generated features (experiment 2) are significantly better than this random sample (τ (434) = −76.71, $p < .05$ and τ (434) = −57.05, $p < .05$ respectively), which further indicates that the prediction results are solid. The comparison of results is given in Fig. 7.1.

To judge the predictive power of each of the features involved, feature ablation for each of the experiments was performed. We compared the performance of the classifier trained on human-annotated 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 results showed that the interaction of the features was dependent on the class to be predicted. For example, grammatical aspect and narrativity seem only to be important for the unbounded class. This finding confirms the results of the multi-factorial analysis carried out in Sect. 4.4, in which the interaction between narrativity and boundedness was a statistically significant factor for predicting the verbal tense in the target language. The verb's POS tags seem to be more informative for the bounded class. However, the adverbs and the infinitives are the features with the most predictive power for both classes. The knowledge about the French verbal tense, the position of the verb (main or subordinate clause) and the verb's children dependencies are less informative than the other features.

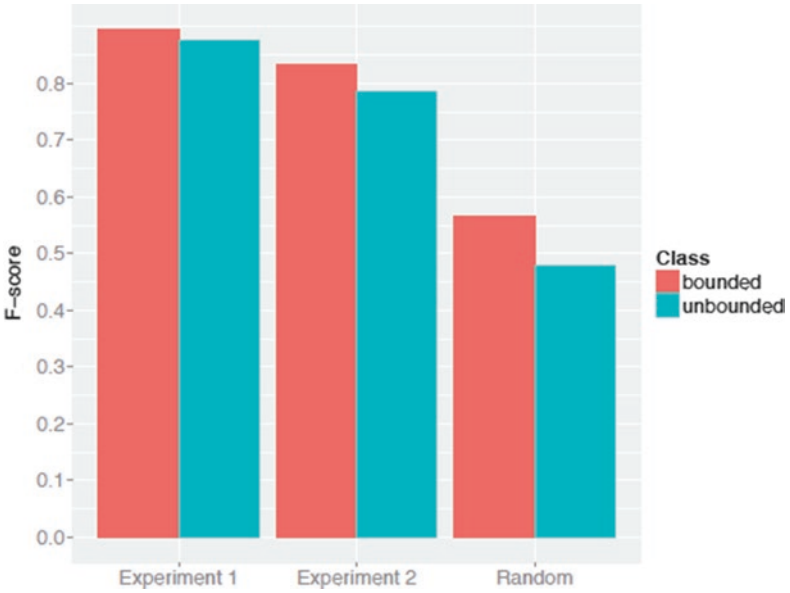

Fig. 7.1 Comparison of results in the three classification experiments

In conclusion, the three NLP experiments described in this section indicated that the [±narrativity] and [±boundedness] features are both identifiable automatically, but that the former is more difficult than the latter. This difference points to the differences in their nature: procedural for the [±narrativity] feature, and conceptual for the [±boundedness] feature. These differences are visible both in both the human and automatic processing of language.

7.2.2 Machine Translation Experiments

The classifiers presented above were built with the purpose of automatically annotating large amounts of data, necessary for the training of SMT systems. Below, I will describe machine translation experiments performed with SMT systems aware of the [±narrativity] feature (Meyer et al. 2013; Meyer 2014) and the [±boundedness] feature (Loáiciga and Grisot 2016).

7.2.2.1 MT Experiments with the [±narrativity] Feature

One key question that arose at this point of the research was how to provide an SMT system with the linguistic information conveyed by the labels given by the classifier. Two methods were tested:

- Concatenation of the label with the Simple Past verb form, considered as a new word to be translated, as in example (625), containing an input sentence for the SMT system in which the concatenation is shown by the '-' symbol.
- Use of factored translation models, which allow for any linguistic annotation to be considered as an additional feature, next to the basic features of the phrase-based models, as in example (626), containing an input sentence for the SMT system, in which the factorization is shown by the 'l' symbol.

To evaluate the gain that the [±narrativity] feature brings to the quality of the translation performed by an SMT system, three systems were built using a 5-gram language model. The first, called the baseline system, is a statistical system trained on plain text input, without verbal labels, as in (624). The second, called the tagged system, is a statistical system using a phrase-based translation model, trained on plain text input containing narrativity labels concatenated on the verb, as in (625). The third, called the factored system, is a statistical system using a factored translation model and trained on texts, where each Simple Past occurrence has a narrativity label whereas all the other words have a |Null label, as in (626), where the Null labels were omitted for legibility (from Meyer 2014, 109).

(624) Baseline SMT: On Wednesday the ČSSD declared the approval of the next year's budget to be a success. The people's party was also satisfied.
(625) Tagged SMT: On Wednesday the ČSSD declared-*Narrative* the approval of the next year's budget to be a success. The people's party was-*Non-narrative* also satisfied.
(626) Factored SMT: On Wednesday the ČSSD declared|*Narrative* the approval of the next year's budget to be a success. The people's party was|*Non-narrative* also satisfied.

To label the SMT data, no manual annotation was used. In the first stage, the actual Simple Past occurrences were identified using the Stanford POS tagger (Toutanova et al. 2003). These tags were replaced by the narrativity labels provided by the MaxEnt classifier, previously built and presented in Sect. 7.2. As pointed out by Meyer (2014), both of the automatic tools used (the POS tagger and the MaxEnt classifier) are prone to errors, which in the end lead to translation errors. However, the challenge was in finding evaluation methods that would allow for the acknowledgment of SMT improvement with respect to the baseline, despite the noisy training and testing data.

As such, for the labelling of the data with the [±narrativity] feature, the MaxEnt classifier described in Sect. 7.2 was used to annotate data for training (in which the system learns the most likely correspondences and re-orders the chunks of words) from the EuroParl corpus (Koehn 2005), containing 321,577 sentences originally written in English and translated into French. Among these sentences, 66,143 instances of Simple Past were identified by the POS tagger used. The classifier labelled 30,452 narrative Simple Past occurrences, and 35,691 non-narrative Simple Past occurrences. For tuning (in which the system trains on a much smaller text in

Table 7.3 Evaluation of SMT systems aware of the [±narrativity] feature

Translation model	BLEU	TER
Baseline	21.4	61.9
Tagged	21.3	61.8
Factored	21.6*	61.7*

order to optimize the language pairs identified in the first stage), the Newstest 2001 tuning set (made available by the Workshop on Machine Translation: www.statmt.org/wmt12) was used, containing 1401 automatically labelled Simple Past instances, of which 807 were narrative and 594 were non-narrative. For testing (in which a new text is handed to the system for translation), the Newstest 2010 data were used, containing 1156 automatically labelled Simple Past instances, of which 621 were narrative and 535 were non-narrative. The SMT system was created using Moses SMT toolkit (Koehn et al. 2007), and applied a 5-gram language model over the entire French component of EuroParl.

The results of the three SMT systems were evaluated using two measures: BLEU and TER. As noted in Sect. 7.1.2, the BLEU score counts the overlap in terms of matching number of words and n-grams between the candidate translation and one or more reference translations. The more matches there are for 4-, 3-, 2- and 1-grams in a candidate translation compared to its reference, the higher the BLEU score. The values of the score range from 0 to 100, where the latter indicates identical translations. The TER measure, for Translation Error Rate (Snover et al. 2006), computes the number of edits (called edit-distance) required to transform a candidate translation into one of its references. The smaller the edit-distance is, the lower the score—thus, the better the translation. Table 7.3 provides the results of the evaluation of the SMT systems in terms of BLEU and TER scores. The factored model improves performance over the baseline by +0.2 BLEU and −0.2 TER (since smaller scores represent better translation), and these differences are shown to be statistically significant at a 95% level of confidence, $p < .05$ according to a t-test (signalled by the * in the table).

Meyer et al. (2013) explain that the lower scores of the tagged model may be due to the sparsity of the data—i.e. verbal forms were altered by concatenation with the narrativity label. As for the small improvement of the factored model, this can be explained by the fact that the narrativity feature improved the translation of the verbal tense alone, and that the translation of the other words in the sentence is unchanged compared to the baseline. So, only a small fraction of the words in the test data are changed, corresponding only to Simple Past occurrences.

A human evaluation of the performance of baseline and factored systems was also performed on the 207 first instances of Simple Past. Bilingual evaluators (English and French) scored the translation by looking at the source sentence and its reference translation from the parallel corpus. The scoring was based on the following criteria: the correctness of the narrativity label, and the improvement of the lexical choice, the choice of verbal tense and the choice of the verb phrase, compared to the baseline system. Human evaluation revealed that the narrativity feature helped

Table 7.4 Human evaluation of verb translations into French, comparing the factored model against the baseline

Criterion	Rating	N.	%	Δ
Labeling	Correct	147	71.0	
	Incorrect	60	29.0	
Verbal tense	Better	35	17.0	
	Same	157	75.8	+9.7
	Worse	15	7.2	
Lexical choice	Better	19	9.2	
	Same	176	85.0	+3.4
	Worse	12	5.8	

the factored system to generate more accurate French verbal tenses in 10% of cases, and to have better lexical choices for verbs in 3.4% of cases, as shown in Table 7.4. The Δ values show the clear improvement of the narrativity-aware factored translation model.

For example, the input English sentence in (627) was translated by the baseline system as in (628), and by the factored system as in (629). The Simple Past *looked* is translated by the baseline system as *considérés* (an infelicitous lexical choice, past participle form, and wrong number agreement), whereas the factored system translates it as *semblait* (a better lexical choice, the Imparfait verbal tense, and correct agreement in number).

(627) Tawa hallae looked|*Non-narrative* like many other carnivorous dinosaurs.
(628) Tawa hallae *considérés* comme de nombreuses autres carnivores dinosaures.
(629) Tawa hallae *semblait* comme de nombreux autres carnivores dinosaures.

Another issue identified by the human evaluation process concerns cases where the factored model performed worse than the baseline system. Some of these cases are due to errors in the POS tagging used to find the Simple Past instances to be labelled. For example, for passive forms of the verb such as *was born*, the auxiliary and the past participle were identified as two separate verbal entities, which were tagged separately: *was* as non-narrative, and *born* as narrative. This introduced noise and errors in the automatic annotation process. Moreover, the factored translation model seems to operate at the local level, despite the pragmatic nature of the [±narrativity] feature. Meyer et al. (2013) suggest that, to widen the context captured by the translation model, one possibility would be to label the entire verb phrase in hierarchical or tree-based syntactical models. Overall, the factored system produces better translations of the Simple Past verb phrase in 9% of cases, compared to the baseline system.

The improvement in translation presented here is important because it points out that it is useful to add pragmatic knowledge about the temporal relations holding between eventualities. However, this numerical value depends on the classifier's performance, which produces reliable but imperfect results (70% correctly labelled

Simple Past occurrences). Recall that the classifier's performance is similar to that of humans, as showed in Experiment 3 for English, and Experiments 8, 9 and 10 for Italian, Romanian and French respectively. According to the HD model of temporal reference, the [±narrativity] feature indicates procedural information which humans cannot easily access by conscious thought. The performance of the classifier indicates that the upper limit found for humans is the same as for machines.

It could be hypothesized that the classifier's better performance would increase the translation quality as well. This was found for the [±boundedness] feature, whose automatic identification accuracy reached an F1 score of around 0.88, as discussed in Sect. 7.2. Below, we will discuss the MT experiments with this feature.

7.2.2.2 MT Experiments with the [±boundedness] Feature

Another series of MT experiments targeted the [±boundedness] feature, in order to assess how much a system enhanced with boundedness knowledge improves the translation of the English Simple Past into French. Phrase-based SMT systems often generate only the most frequent translation possibility, the Passé Composé, as the default. The goal of these SMT experiments was to provide a system with bounded/unbounded labels in order to boost the other three tenses, improving the verbal tense translation choice.

Given that there is no data set annotated with this aspectual information sufficiently large to train an SMT system, the corpus was automatically annotated with the [±boundedness] feature, using the classifier described in Sect. 7.2. The data were taken from the MultiUN corpus, a corpus of translated documents from the United Nations, provided by Eisele and Chen (2010). All English Simple Past occurrences are identified and labelled as either bounded or unbounded automatically. The training data consisted of 350,000 sentences (134,421 Simple Past instances), the tuning data consisted of 3000 sentences (1058 Simple Past instances), and the testing data consisted of 2500 sentences (1275 Simple Past instances).

Loáiciga and Grisot (2016) report that the Moses Toolkit (Koehn et al. 2007) was used to build two systems: a baseline without boundedness labels, and an aspect-aware system with such labels. Both systems are phrase-based models with identical composition, trained, tuned and tested on the data described above, and use a 3-gram language model built using KenLM (Heafield 2011) and trained on over ten million sentences of French monolingual data, taken from the 2015 Workshop on Machine Translation (Bojar et al. 2015).

As with the MT experiments on the [±narrativity] feature, the boundedness labels are combined with the SMT system using a factored model (Koehn and Hoang, 2007). Instead of the standard text, the system is trained on annotated text of the form shown in (630). The example shows an input sentence labelled by the classifier as follows: the verb receives an *unbounded* label, whereas all other words from the sentence receive a Null label.

Table 7.5 Evaluation of SMT systems aware of lexical aspect

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

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

As in the case of the narrativity classifier, no single factor entirely determines the translation of a verb—i.e. there is no exact correspondence between a label and a verbal tense in French. For instance, a *bounded* label does not necessarily lead to a translation into French by a Passé Composé. Instead, various factors are considered when estimating the translation probabilities computed over the entire parallel corpus.

The performances of the two translation systems were evaluated with the BLEU measure, computed across all the sentences in the test set, as well as the sentences containing a Simple Past only. The results provided in Table 7.5 indicate that the factored system using the lexical aspect labels led to an increase of 0.98 points. When computing the BLEU score for the sentences with Simple Past verb phrases only, there was a difference of 1.58 points. These scores reflect an improvement in the quality of the translation of Simple Past occurrences. On the one hand, these increments suggest that the method does not degrade the general translation quality of all the other words in the sentence; on the other hand, they suggest that it does not change the Simple Past translations already estimated to be adequate by the baseline model. Loáiciga and Grisot (2016) points out the importance of this result, given that the boundedness-aware system only targets Simple Past occurrences and not all the words in the sentence.

This score may be further analysed using the bootstrap resampling significance test (Koehn 2004). This test estimates the difference in performance of one SMT system in comparison to another. The output of the lexical aspect-aware translation system was compared to the output of the baseline SMT, in terms of the translation of the same 300 sentences. For each sentence in each sample, a BLEU score was computed. The analysis of the 300 BLEU scores showed that, in 50% of the sentences, the BLEU scores of the aspect-aware system are higher than the scores of the baseline system. In other words, at least one English Simple Past verb was better translated by the aspect-aware system than by the baseline system.

Automatic metrics and statistical tests do not give any further indications of the particular qualitative differences in the translation of verbal tenses between the outputs. To overcome this, a human evaluation of the performances of the two systems and of the performance of the classifier was carried out on 200 randomly selected instances of the Simple Past. The classifier correctly identified the Simple Past instances in 91% of cases, and correctly annotated them as bounded or unbounded situations in 65% of cases. In general, the bounded class seems more difficult to

predict than the unbounded class. The manual evaluation also revealed that several verbs which usually express one-time events, like *ask*, *request*, *result*, *adopt*, *add* or *call*, were treated as though they had a duration which is much less common. Finally, several instances of the same verb appeared repeatedly, and the same classification error was thus repeated: for example, *was* labelled as bounded.

For the factored SMT system, compared to the baseline system, human evaluation indicated a better translation of Simple Past instances into French in 25% of cases, a similar translation in 54% of cases, and a degraded translation in 21% of cases. The cases of similar translation can be explained by the fact that the baseline system itself performed well, since it provides Passé Composé translations by default, and the distribution of the verbal tenses used in the translation into the target language is highly skewed in favour of the Passé Composé. Therefore, the improved cases are those where an Imparfait was used in the reference, and the aspect-aware system correctly translated a Simple Past by an Imparfait. For example, the input English sentence in (631) was translated as (632) by the baseline system, as (633) by the factored system aware of boundedness, and as (634) by a professional translator, the reference translation coming from the parallel corpus.

(631) The vice-chairman of the ODS, Petr Nečas *said* that the concept of an interim government supported by the ČSSD, ODS, and Green Party, *was* evidently no longer working.

(632) Le vice-président, de l'ODS Petr Nečas, *dit* que le concept d'un gouvernement intérimaire soutenu par les ČSSD, ODS, et parti vert, *a* apparemment aucune *fonctionne* plus.

(633) Le vice-président, de l'ODS Petr Nečas, *a déclaré* que le concept d'un gouvernement intérimaire soutenu par les ČSSD, ODS et aux verts, *était* manifestement, de ne plus travailler.

(634) Le porte-parole de l'ODS Petr Nečas *a déclaré* que l'idée d'un cabinet administratif soutenu par le ČSSD, l'ODS et le Parti des verts ne *fonctionnait* manifestement plus.

The first Simple Past, *said*, was labelled by the classifier as *bounded*, and the second Simple Past, *was*, as *unbounded*. Both verbal tenses were translated with a Présent by the baseline system. The factored model instead produced the same verbal tenses as the reference: Passé Composé for the first Simple Past and Imparfait for the second. The 21% of examples which were degraded were possible outcomes, given that these translations are possible outcomes of the factored model's non-deterministic disposition. This result is also directly linked to the results of the bounded/unbounded labelling: correct labels entail twice as many improved translations.

Overall, the factored system produces better translations than the baseline. An improvement can also be observed if the two factored systems (i.e. one aware of temporal information, and the other aware of lexical aspect) are compared. The aspect-aware SMT system produced better translations than the narrativity-aware SMT system (15%). This is mainly due to the better performance of the classifier

producing boundedness labels than the classifier producing narrativity labels. The second reason is the better identification of correct instances of the Simple Past. This was due to the use of the POS tagger, improved with a series of rules. Recently, other methods have been suggested, such as direct document-level translation (Hardmeier et al. 2012; Hardmeier 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 have done here. Both methods have proved their efficiency in comparison to a baseline system.

To conclude, I would like to point out the importance of the granularity of the linguistic features. To be usable, linguistic features must be medium-coarse grained. In other words, features which are too fine-grained are either insufficiently capable of explaining the variation in the data, or they are not implementable. For example, the mixed statistical model based on the manually annotated corpus of 435 sentences (cf. Sect. 4.4) shows that the French verbal tense in the target language is significantly determined by the interaction between the narrativity status and the lexical aspect of English verbs. This theoretical insight is unfortunately very difficult to model in NLP, and to apply in SMT. This is an important issue to be investigated in further research. In Loáiciga and Grisot (2016), we 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 class task (+narrative +bounded, +narrative −unbounded, −narrative +bounded, and −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 thus two different factors in the factored model. It should be tested whether or not diluting the information in such a way would still add knowledge to the system, since the distributions may result in insufficient data.

7.3 Summary

My aim in this chapter was to show that the role of Tense and Aktionsart in language processing was also validated from NLP and MT perspectives. Research on the automatic processing of temporal reference has focused in the past few years on issues such as event ordering (events relative to one another), time stamping (i.e the temporal anchoring of a situation) and generation of words expressing temporal relations for individual languages, usually for English. Some of the most influential studies are those demonstrating that Tense is an anaphoric category (Partee 1973, 1984) and Webber (1988), as well as exploring the role played by Aktionsart in determining discourse structure (Dowty 1979, 1986; Moens and Steedman 1987, 1988; Steedman 1997; Passonneau 1988). As for tenseless languages, such as Mandarin Chinese, it was shown that the most relevant temporal indicators are temporal adverbials, aspectual markers, special verbs and prepositions (Li et al. 2001,

2004). In the field of MT, Ye et al. (2006) used telicity and event ordering to generate verbal tenses when translating from Chinese to English.

Meyer et al. (2013), Meyer (2014) and Loáiciga and Grisot (2016) have shown that these two properties—operationalized as the [±narrativity] and [±boundedness] features—also significantly improve the results of SMT systems when the source and the target languages are tensed languages. In particular, they have shown that these two features can be automatically identified in raw data by classifiers which have been previously trained on human-annotated data. When these classifiers are integrated into an SMT system, the translation is better than that of a baseline system, in terms of lexical choices and inflection choices for verbs.

Not only can the medium-coarse grained features proposed—i.e. [±narrativity] and [±boundedness]—be implemented successfully, their implementation in NLP and application to MT produced significant improvements in the results of the automatic systems. As such, these ameliorations provide an indirect but solid empirical validation of the theoretical model proposed in Chap. 5.