Toward Temporally-aware MT: Can Information Extraction Help Preserve Temporal Interpretation?

Taylor Cassidy taylor.cassidy.civ@mail.mil
Army Research Laboratory, Adelphi, MD 20783, USA
Jamal Laoudi jamal.laoudi.ctr@mail.mil
ARTI, Fairfax, VA 22030
Clare Voss clare.r.voss.civ@mail.mil
Army Research Laboratory, Adelphi, MD 20783, USA

Abstract

Users of MT systems often need to glean information about the world from foreign language texts for specific tasks, such as documenting how events, as mentioned in those texts, fit on a time line. Current systems have not been systematically evaluated for their adequacy in preserving *temporal interpretation*, i.e., the set of temporal relations that a reader naturally takes to hold among the states, events, and time expressions mentioned in the text, as well as the intrinsic temporal properties of each, although some MT research has focused on exploiting linguistic mechanisms, such as verbal tense or aspectual markers to convey temporal information. We describe ongoing work to develop a method for (i) building parallel TimeBanks with annotated temporal interpretation on parallel texts, (ii) leveraging these resources to train and evaluate the emerging class of *temporal interpretation extraction systems* on new languages, and (iii) developing *time-aware MT systems* that aim to preserve the temporal interpretation of source language text in their target language outputs. We present our approach and results from our exploratory analyses into the preservation of temporal interpretation in Arabic-English MT, and propose shared tasks to bring together research in information extraction and machine translation, geared toward building time-aware MT..

Users of MT systems often need to be able to glean information about the world from foreign language texts for specific tasks, such as documenting their understanding of how events, as mentioned in those texts, fit on a time line. While task-based metrics have evaluated the extent to which MT preserved who, when, and where information (Voss and Tate, 2006) or information required to pass language proficiency tests (Jones et al., 2005; Matsuzaki et al., 2015), current systems have not been systematically evaluated for their adequacy in preserving *temporal interpretation*, i.e., the set of temporal relations that a reader naturally takes to hold among the states, events, and time expressions mentioned in the text, as well as the intrinsic temporal properties of each.

Some MT research has focused on exploiting linguistic mechanisms, such as verbal tense or aspectual markers, when available in text to convey specific forms of temporal information. MT systems incorporating this research address the challenge of preserving temporal content narrowly, e.g., selecting the correct target language tense for each source language verb, or selecting the correct sense to translate temporal discourse connectives. However different languages rely on a much wider range of explicit temporally-significant linguistic mechanisms to convey underlying temporal content, including tense, aspect, function words, discourse connectives, syntactic relations, idiomatic expressions.[1] These mechanisms are manifested asymmetrically across languages; as a result, reference translations may use different mechanisms compared with those in the source.[2] Thus proper selection and use of these mechanisms is non-trivial. The MT challenge of preserving temporal interpretation from source language to the target language output goes beyond current approaches and necessarily subsumes working with many forms of temporal information.[3]

A growing body of computational research now studies temporal interpretation in text, having initially emerged to support systems performing tasks such as information extraction and knowledge base construction, and thus has generally taken place outside the MT research community. This research has become multilingual and a variety of corpora in many languages, including parallel corpora (Forăscu and Tufiş, 2012), have been annotated using the TimeML annotation scheme (Pustejovsky et al., 2003a). In addition, temporal interpretation algorithms implemented within extraction systems have successfully made use of a variety of features drawn from the linguistic temporal mechanisms listed above (UzZaman et al., 2013; Bethard et al., 2016), though no one feature type stands out as dominant. Despite the fact that such annotation frameworks and automatic extraction algorithms for temporal interpretation exist, and translation of certain temporal linguistic mechanisms has been improved, little has been written about explicitly preserving temporal interpretation in MT. Simply put, MT engines are not built to be fully temporally-aware.

This paper describes ongoing work to develop a method for (i) building parallel TimeBanks with annotated temporal interpretation on parallel texts, (ii) leveraging these resources to train and evaluate the emerging class of *temporal interpretation extraction systems* on new languages, and (iii) developing *time-aware MT systems* that aim to preserve the temporal interpretation of source language text in their target language outputs. We present our approach and results from our exploratory analyses into the preservation of temporal interpretation in Arabic-English MT, and conclude by proposing shared tasks to bring together research in information extraction and machine translation, geared toward building time-aware MT.

2 Temporal Interpretation and Its Preservation Across Languages

2.1 Definition of Temporal Interpretation

Similar to Katz and Arosio (2001)'s "radically simplified semantic formalism," we start with the notion that the temporal interpretation of a text is *the set of temporal relations that a reader naturally takes to hold among the states, events, and time expressions mentioned in the text, as well as the intrinsic temporal properties of each.* Intuitively, the temporal interpretation provides sufficient information to answer questions about when events occur and states obtain to the level of precision and certainty intended by the text's author. Temporal interpretation is independent of a text's accuracy with

[1] E.g., "when women go to the farm" can mean during the hours 8-9am in Babungo Schaub (1985).

[2] Indeed, this asymmetric usage may not be optional. For example, Chinese has no clear analogue to English grammatical verb tense.

[3] We distinguish *information*, as processed raw text data, from *interpretation*, as a process by which information is grounded or annotated for inclusion, for example, in a knowledge base or a time line. (The processing in either case may be manual or automated.) As explained further in section 2, for us, temporal information is a narrow term that refers to explicit surface mentions with temporal content, while temporal interpretation is a broader, more inclusive term for a process, or the result of a process, that starts with temporal information, but also allows for inference to derive implicit temporal entities and relations using world knowledge, such as date-time arithmetic.

respect to what transpires in the actual world, and for temporal relations left vague in a text, their interpretation may vary among readers due to the unique perspectives and prior knowledge that they each bring to understanding the text.

Our working definition of the temporal interpretation of a text is inspired by the TimeML annotation scheme, as used by Pustejovsky et al. (2003b). Here we focus only narrowly on core aspects of temporal interpretation to give the reader a sense for what interpretation means procedurally, first in terms of what information is annotated in text and second, how the annotated information then is connected to specific concepts of time.[4] We use the label SL to stand for a source language text and $I(SL)$ for the temporal interpretation over that text.[5] We define that interpretation as a five-tuple, $\langle SL_e, SL_t, \tau_{SL}, \sigma_{SL}, r_{SL}\rangle$. First, two sets designate what information is annotated: the set SL_e consists of the events and states (henceforth shortened collectively to "events") mentioned in SL and the set SL_t contains the time expressions mentioned in SL. The elements of their union are referred to collectively as *temporal entities*. Second, the remaining items in the five-tuple are functions that designate how these entities are connected to concepts of time. The function τ_{SL} maps mentioned time expressions, SL_t, into actual times, as captured in T, the set of all possible time values (e.g., as identified in ISO-8601). The function σ_{SL} maps mentioned events, SL_e, into a space of semantic property values, such as *class*, *polarity*, *modality*, *tense*, etc.[6] The function r_{SL} maps pairs of temporal entities into temporal interval relations, elements in set S (e.g., as in Allen's interval relations).

Several shared task workshops in the past decade have tackled automating the temporal interpretation of text. To evaluate the algorithms in computational systems built for this task, the TempEval and Clinical TempEval tracks of previous SemEval workshops used, as their ground truth, manually annotated corpora such as TimeBank (Pustejovsky et al., 2003b) and THYME (Styler IV et al., 2014). Both corpora use a version of TimeML annotation schema, with guidelines for identifying and annotating event and time expression words or phrases (defined above as SL_e and SL_t). Each *temporal entity* corresponding to an event or time expression is conceptualized as having a *temporal extent* that is either an interval or a set of intervals.[7] Events in TimeML are also labeled with semantic properties such as *class, polarity, modality, tense*, etc. At the current time, inter-annotator agreement (IAA) rates however on the manual task of annotating texts for temporal interpretation have varied considerably depending on the particular setting and they tend to be lower than for similar text annotation tasks. For example, TimeML relations generally achieve Kappa scores between .4-.8, while PropBank annotations for argument roles achieve higher .91-.96 Kappa (Palmer et al., 2005).

Nonetheless even with such varied IAA results, researchers have built computational systems to perform temporal interpretation extraction[8] with supervised machine learn-

[4] We later extend this to less obvious sources of temporal information, such as definiteness.

[5] Similarly TL and $I(TL)$ stand for a target language text and its temporal interpretation.

[6] That space is $\times_i^N S_i$, the cross product of N sets of event semantic property values, S_i.

[7] Other details of TimeML's interval annotation include the following. End points may be precise dates and times, or as less specific values from a pre-defined set, e.g. "morning". Intervals may or may not be anchored to an actual time line. Time expressions are assigned their extents directly via a Timex3 tag. When an event and its location in time are determined, that pair of temporal entities is tagged with a TLINK characterizes their relationship (e.g. Before, Overlap). TLINKs are also used to relate events in time relative to other events.

[8] We refer information extraction (IE) systems that focus explicitly on incorporating temporal interpretation into its annotations as *temporal interpretation systems* or *temporal extraction systems.*

ing algorithms to identify events, time expressions, and their semantic properties. The systems work with features for word and character n-grams, POS tags, information from lexical ontologies, and distributional semantic vectors. For labeling higher-order temporal entity pairs with interval relations, systems make use of additional types of linguistic knowledge such as syntactic context from the dependency path between temporal entities. For the specific subtask of time expression normalization, the highest-performing systems have generally been rule-based (Strötgen and Gertz, 2010; Chang and Manning, 2012). Researchers have had less success in modeling implicit world knowledge in temporal extraction systems. For example, Mirza and Tonelli (2014) observed that adding a feature encoding typical event duration to an SVM classifier decreased system accuracy. Unsurprisingly, annotation guidelines differ on whether such inferences are allowed. The TempEval corpora annotation guidelines encourage the use of world knowledge (Verhagen et al., 2009), while the still-evolving Richer Event Description (RED) guidelines prohibit its use.[9]

To date, manual temporal interpretation annotation is carried out monolingually by native speakers of the language of the text being annotated, who are trained with language-specific guidelines. And so temporal interpretation systems are also constructed independently in separate languages based on those resources. Schematically, we illustrate this situation with separate rows in figure 1a[10].

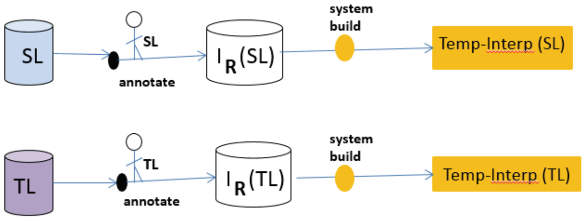

Figure 1a: Annotation of Temporal Interpretation in Texts for System Construction: Monolingual workflows here are independent for source & target language (SL & TL).

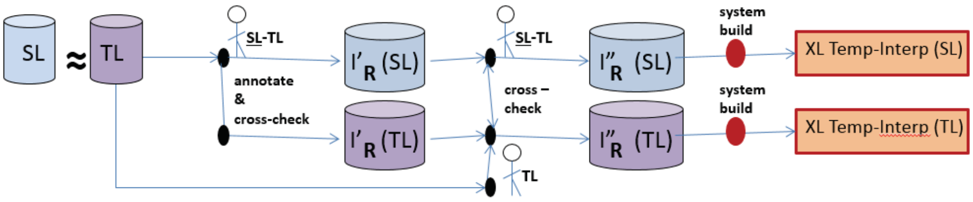

Figure 1b: Annotation of Temporal Interpretation in Parallel SL and TL Texts for System Construction: Cross-lingual workflows here are interdependent by establishing semantic correspondence of temporal annotations in SL & TL prior to system builds.

[9] For example, given the text "We diagnosed her cancer last week", TempEval guidelines would permit annotating that the cancer preceded the diagnosis, whereas RED guidelines would not. (see https://github.com/timjogorman/RicherEventDescription.)

[10] In our figures, cylinders represent datasets, circles represent human or automated processes, and colored rectangles represent software systems.

2.2 Preservation of Temporal Interpretation

Given that established methods now exist for annotating temporal interpretation in text and that computational systems can extract temporal information, we now ask, how might these monolingual methods, corpora, and systems be leveraged to work cross-lingually to *preserve* temporal content from the source language in machine translation to the target language?

When evaluating the translation of a text from one language to another, it is natural to ask whether the meaning of the text is fully preserved by the translation. Here we focus on preservation of the text's temporal interpretation only. Consider two texts: an SL and its translation TL. As show in figure 1a, let $I_R(SL)$ and $I_R(TL)$ be their temporal interpretations as derived independently by a native speaker of each language.[11] We say that the interpretations are *semantically equivalent* when all their identified core components are *semantically equivalent*, i.e., their identified temporal entities correspond cross-lingually (SL_e with TL_e and SL_t with TL_t) and the values of temporal grounding functions (τ_{SL}, σ_{SL}, and r_{SL}) over SL identified entities or entity pairs in their domains, correspond cross-lingually with function values (τ_{TL}, σ_{TL}, and r_{TL}) over TL identified entities or entity pairs in their domains. When $I(SL)$ and $I(TL)$ are semantically equivalent, we say that the translation TL *strongly preserves* the interpretation $I(SL)$.

This formal notation is of course a conceptual-level abstraction, removed from the realities of actual cross-language divergences in how and where temporal information is expressed. Information that is explicitly lexical in a sentence in one language may be grammaticalized in its translation in another language, and left implicit in another language. These divergences complicate the detective work of identifying the temporal content that is preserved in actual translation by humans and machine translation. Intuitively, we would like to say that a translation of SL to TL preserves temporal content to the extent that native speakers of each language independently arrive at the same temporal interpretation. However, even parallel texts may yield distinct temporal interpretations, and so we have begun experimenting with a process for converging those interpretations as part of the annotating process over parallel texts.[12]

Operationally, we consider the most likely scenario for such preservation in corpus construction, when annotators seek to interpret temporal information in parallel texts. We assume both a bilingual SL-TL annotator who is native in the SL, shown schematically in figure 1b as stick person labeled SL-TL with SL underlined, and a monolingual TL speaker, also shown as a stick person with label TL only. The bilingual annotator can read given parallel SL and reference TL texts, and develop annotations by cross-checking texts and their temporal interpretations in tandem. We designate their initial interpretations I'_R, where the single quote indicates the first pass and the subscript R indicates reference interpretation.[13] The bilingual aims in their annotation for semantic correspondence of $I'_R(SL) = I'_R(TL)$. For quality control, the native TL speaker can read TL and $I'_R(TL)$, and then cross-check the TL annotations for semantic correspondence to the annotations only in $I'_R(SL)$, by conducting a systematic review of temporal entities and their relations in both texts with the bilingual annotator. Since first-pass interpretations and the reference translation itself may be changed in the review, we adopt the second-pass results, $I"_R(SL)$ and $I"_R(TL)$, as reference interpretations for

[11] The subscript R indicates these are human reference interpretations.

[12] Our initial efforts in developing this process are described in section 3.

[13] The quote and subscript will later serve to distinguish this from, respectively, subsequent second pass with double quotes $I"_R$ and automated interpretations with subscripts to identify the system.

system builds and evaluations.

We also extend the notion of preserving temporal information for the purpose of evaluating machine translation output, $"TL"$.[14] When an MT user who knows they are reading MT output and uses that knowledge in drawing inferences, then creates a temporal interpretation of $"TL"$, we denote their interpretation $\hat{I}("TL")$. Furthermore, we say the MT engine has *weakly preserved* the source interpretation when the MT output has translation errors in conveying that original temporal content, but the user can overcome that information with their background knowledge, as shown when subsequent evaluation of the MT user's $\hat{I}("TL")$ shows it to be semantically equivalent to the reference $I''(TL)$.

2.3 Time-Aware Machine Translation: Current State of the Art

Our long-term research goal is to develop time-aware machine translation systems that preserve temporal interpretation from the source text in the target language output. The most relevant research efforts have aimed to correctly translate specific types of linguistic mechanisms that speakers use to convey temporal interpretation. These efforts therefore indirectly aim to preserve temporal interpretation. Access to tense (Klavans and Chodorow, 1992), lexical aspect and temporal connectives (Dorr and Gaasterland, 2002), have been shown to help lexical choice, critical for all MT systems. For statistical systems, automatic prediction of target language tense based on source language verbs has been a popular task, especially for Chinese-to-English translation due to the lack of overt tense on Chinese verbs (Olsen et al., 2001; Ye et al., 2006; Baran, 2013; Ge et al., 2015; Loaiciga et al., 2014).

There have been few efforts to integrate linguistic temporal mechanisms from source text directly into a statistical MT system, and evaluate its impact on MT performance. Meyer et al. (2013) used a factored translation model that included a binary narrativity tag on each English source verbs in the simple past tense as a feature to improve choice of French output tense. Loaiciga et al. (2014) extended this model by using a supervised machine learning classifier to further tag each English source text verb with one of nine possible French verb tenses. It's worth noting that one of the features used by English verb tagger was derived from event tense, aspect, and class for event pairs as labeled by a temporal interpretation extractor. They were able to achieved 10% improvement on tense translation. Gong et al. (2012) re-score translation hypotheses during decoding time using (1) English tense labels automatically assigned to Chinese source-language verbs, and (2) a tense n-gram language model that models the probability of a given sequence of tenses in English text. Meyer et al. (2015) use a factored language model to leverage discourse connective marker types. They train a supervised classifier to predict Penn Discourse Treebank style tags on discourse connectives, whose values include time-relevant values such as temporal, temporal-durative, temporal-punctual, temporal-contrast, and temporal-causal. This work uses similar features to Loaiciga et al. (2014).

While it is reasonable to expect improving translation of verbal tense or discourse connectives in particular will increase the likelihood that temporal ordering will be preserved, we are eager to develop a broader approach to improving temporal interpretation preservation in MT. We will return to these questions pertinent to this broader approach after the exploratory analysis section: (i) What resources are needed to assess the extent to which current MT systems preserve temporal information? (ii) How might temporal interpretation extraction systems be integrated into the workflow for building

[14] The quotation marks distinguish MT output as $``TL''$ from reference translation text labeled TL, to remind readers that the accuracy and fluency of MT output is rarely equivalent to its reference TL.

MT engines, and how will this impact the quality of MT output?

3 Exploratory Analysis

This section documents our ongoing effort to determine how well current MT systems preserve temporal interpretation. Our procedure is depicted in Figure 2. The results of the procedure are a reference translation that strongly preserves the temporal interpretation of the source, MT output for each source sentence, and an evaluation of whether MT output weakly preserves the source's temporal interpretation.

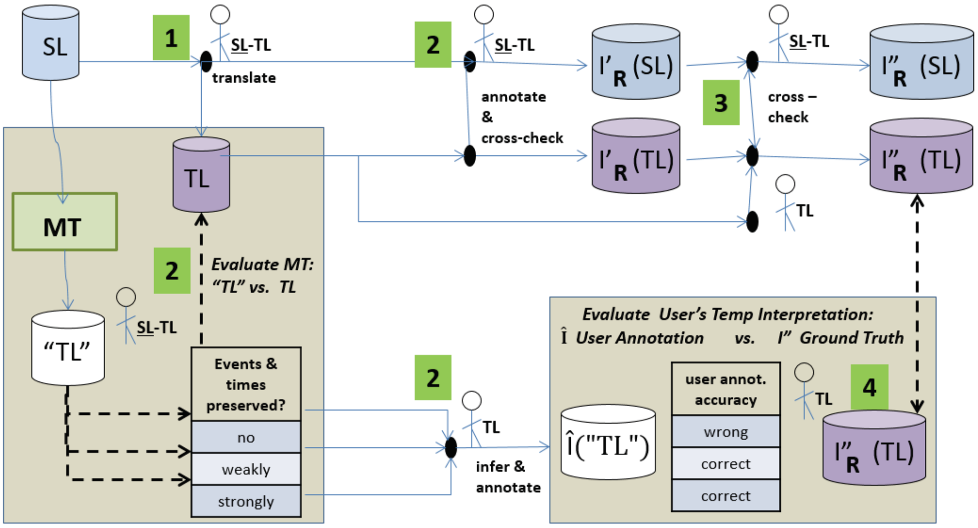

Figure 2: Procedure for Exploratory Analysis.

This procedure requires two participants, a bilingual source language native speaker fluent in the target language ($\mathrm{P}_{\underline{\mathrm{SL}}\text{-}\mathrm{TL}}$), and a native speaker of the target language (P_{TL}).

In step 1 $\mathrm{P}_{\underline{\mathrm{SL}}\text{-}\mathrm{TL}}$ translates SL text to obtain TL aiming to preserve temporal interpretation, while MT translates SL to obtain "TL". These translations are independent of each other and so can be done during this same step. In the step 2, $\mathrm{P}_{\underline{\mathrm{SL}}\text{-}\mathrm{TL}}$ derives a temporal interpretation for both SL and TL, $I'_R(SL)$ and $I'_R(SL)$. In addition, $\mathrm{P}_{\underline{\mathrm{SL}}\text{-}\mathrm{TL}}$ identifies the events and time expressions in "TL" corresponding to those identified in the Arabic source, noting where this temporal information is strongly, weakly, or not at all preserved. Meanwhile, P_{TL} independently constructs $\hat{I}(\text{“}TL\text{”})$ without any access to the reference translation.[15] In step 3, $\mathrm{P}_{\underline{\mathrm{SL}}\text{-}\mathrm{TL}}$ and P_{TL} work together on two tasks. P_{TL} provides target language expertise to help $\mathrm{P}_{\underline{\mathrm{SL}}\text{-}\mathrm{TL}}$ alleviate inconsistencies between $I'_R(TL)$ and $I'_R(SL)$, yielding revised interpretations $I''_R(TL)$ and $I''_R(SL)$ and possibly revisions to TL text. $\mathrm{P}_{\underline{\mathrm{SL}}\text{-}\mathrm{TL}}$ works with P_{TL}, who can also consult the reference translation TL during this step of cross-checking. In step 4, after the cross-checking is complete, P_{TL} is then able to assess the extent to which $\hat{I}(\text{“}TL\text{”})$ is equivalent to $I''_R(TL)$, assessing whether MT users can overcome MT output errors, and so allow

[15] A variety of factors bear on one's ability to determine $\hat{I}(\text{“}TL\text{”})$, and the methodological choices made in doing so, such as: (i) knowledge about the MT algorithm (e.g. statistical vs. rule-based),(ii) experience reading MT output, both in general and from the MT engine in question, (iii) knowledge of SL and access to the source text, (iv) experience manually translating the source to TL.

for the possibility that some preservation, what we are calling weak preservation, was indeed achieved by the MT engine.

Our analysis thus far consisted of a shallow pass through the procedure described above. No annotation guidelines were strictly enforced, but we roughly adhered to TimeML guidelines for identifying events and time expressions. Temporal relations were not explicitly annotated; however, both participants were previously trained to perform the temporal relation annotation task, and holistically developed informal temporal interpretations. Rather than compute quantitative results, we identified representative cases illustrating phenomena relevant to preservation of temporal interpretation. Selected results are reported in section 4.

The participants were authors of this paper. P_{TL} is a native English speaker, and $P_{\underline{SL}\text{-}TL}$ is a native Arabic speaker and fluent in English, with machine-aided translation experience. We used 10 documents from DARPA's GALE dataset and several manually translated recent sports news documents.

4 Phenomena Pertaining to Temporal Interpretation Preservation

We encountered a variety of phenomena that lead to a failure to preserve temporal information. However, instances where MT output fail even to weakly preserve the Arabic temporal interpretation are of particular concern given that we are driven primarily by practical usage of MT. In this section we discuss instances where one of the following was the case: (i) English MT output failed to weakly preserve Arabic temporal interpretation (thus, strong preservation failed as well); or, (ii) English MT output failed to strongly preserve Arabic temporal interpretation, but did weakly preserve it, in spite of the fact that elements of temporal significance were incorrectly translated.

Incorrect translation of events: Perhaps the most obvious source of error is when events and time expressions themselves are not translated correctly. In some cases, an Arabic event word that should have been translated was transliterated, or vice versa. In other cases the incorrect English word in MT output can lead the user to unintended interpretation. Table 1 provides two representative examples where temporal interpretation is not preserved

Poor time expression interpretation: Temporal expressions often serve as *temporal containers* with respect to which many events can be related. Rather than annotate every single pair of events, annotators can effectively specify a great deal of a text's temporal interpretation by relating each event to a few key time expressions. Thus, assigning the wrong temporal extent can greatly distort temporal interpretation. We found instances where time expressions rely on the semantics of certain verbs for their correct interpretation. Consider the temporal interpretation of "year" in the following example:

- وقال "لدينا فريق رائع، جعلنا نستمتع في كل مباراة. من المؤكد **أننا بصدد عاما صعبا للغاية**" ،ولكننا مستعدون"

- Ref: He said: "we have an amazing team that allowed us to enjoy every game. Surely **we are embarking on a very difficult year**, but we are ready".

- MT: He said, "We have a great team, made us enjoy every game. Sure, we **are in a very difficult year**, but we are ready."

In the reference translation, "year" is a time period that begins in the future or has just begun speech time. In the MT, by virtue of losing the verb "embarking", we

Arabic	MT Output	Reference	Explanation
المباحثاث تناولت الوسائل الكفيلة بتعزيز التعاون	**Alambagesat** dealt with ... enhancing cooperation	The **discussions** were concerned with ways to enhance cooperation ...	The Arabic word المباحثات was misspelled as المباحثاث, which lead MT to transliterate to “Alambagesat” instead of translating to “discussion”.
تطرق رئيس نادي برشلونة الإسباني، جوزيب ماريا بارتوميو، إلى الحالة الاقتصادية القوية...اليوم	**Turning president** of FC Barcelona, Josep Maria Bartomeu, the strong economic situation ... today	The **president** of the Spanish club Barcelona, Jossepi Maria Bartomero, **addressed** the current strong financial status ... today ...	“addressed” mistranslated as “Turning” could lead to interpretation that the presidency is just beginning today

Table 1: Examples of Incorrect Event Translations with Explanations.

lose the interpretation that the year is just beginning or about to begin is not preserved (however, “we are ready” does suggest this interpretation).

We encountered instances where noun phrase definiteness affects the interpretation of time expressions. Consider the following example:

- وقال في هذا الصدد "نريد غلق بعض الملفات دون تعجل. سندافع عن اللاعبين كلما اقتضت الحاجة"، مشيدا **بالعام الذي "لا ينسى"** على المستوى الرياضي بعد الأربعة ألقاب التي حققها الفريق".
- Ref: He stated about this topic that “we want to close these cases in a timely fashion. We will defend the players as needed,” praising **this “unforgettable” year** at the competitive level after the four trophies the team has won.
- MT: He explained, ”We enjoyed it a lot and managed to destroy the attacking triangle of all the numbers, we’ve lived unforgettable **years** and we have made four titles reflect a lot of personal competitive team.”

Here, the definite singular demonstrative determiner “this” in the reference translation indicates a single year, where the time of speech is at or near the end of that year. The source text uses a definite singular determiner. The indefinite plural “years” in the MT output, however, indicates multiple years in the past, not necessarily contiguous, leaving whether the current year is part of that set of years unspecified. Thus, the source text temporal interpretation is not preserved.

Missing or Mistranslated Function Words in MT Output: A single missing word can dramatically alter the meaning of a sentence, including its temporal interpretation. In the following case the missing word is the conjunction “and”, though we observed missing and mistranslated prepositions as well.

- وتاتي **زيارة**$_{e1}$ ساركوزي الى المغرب **بعد اربعة ايام**$_{e2}$ من **زيارة** رئيس مكتب التحقيقات الفدرالي الاميركي (اف بي آي) روبرت اس مولر الى المغرب **وتسليم**$_{e3}$ الولايات المتحدة ثلاثة معتقلين مغربيين من معتقل قاعدة غوانتانامو الى المغرب

- Ref: Sarkozy's **visit**$_{e1}$ to Morocco comes **four days** after the **visit**$_{e2}$ to Morocco of FBI Chief Robert S. Mueller, **and** the **handing over**$_{e3}$ of three Moroccan prisoners from the Guantanamo detention camp to Morocco by the United States.

- MT: Sarkozy's **visit**$_{e1}$ to Morocco, **four days** after the **visit**$_{e2}$ of the US Federal Bureau of Investigation (FBI) Robert S. Mueller to Morocco, The United States **delivered**$_{e3}$ three Moroccan detainees from Guantanamo Bay base to Morocco.

Both reference translation and MT output show that e1 occurred after e2 (with four days separating the two events). The presence of "and" in the reference indicates that e3 also occurred four days after e2 and an equivalent word plays the same role in the source text. However, the relative ordering of e2 and e3 is not so clear in the MT output. It seems likely that a word or phrase is missing between the comma and "The United States" in the MT output, but its impossible to tell if it should be "and", or "just before", or something else. The missing main verb "comes" adds additional complexity. The source temporal interpretation is therefore not preserved.

Phrases are Incorrectly Ordered: We found instances where incorrectly ordered phrases significantly impacts temporal interpretation.

- MT: Sarkozy's **visit**$_{e1}$ to Morocco after **four days**$_{t1}$ of **visiting**$_{e2}$ Chairman of US Federal Bureau of Investigation (FBI) Robert Mueller Vegas to Morocco and **extradition**$_{e3}$ to The United States three Moroccan prisoners from Guantanamo to Morocco.

The relative order of "after" and "four days" is switched, resulting in the possible interpretation that "four days" describes the length of the "visit" (e2) as opposed to the time elapsed between e1 and e2. Here, a single change in word order leads the reader to favor a syntactic analysis where "four days of visiting" is a linguistically motivated phrase. While this interpretation leaves the existence of the second "to Morocco" unexplained, recovering the fact that "Robert Mueller" is the agent of event e2 and that "to Morocco" describes the location of that visit is quite difficult; thus, temporal interpretation is not preserved.

Tense and Aspect: Tense and aspect are generally considered important cue to temporal interpretation. Most machine learning and rule based temporal information extraction systems use tense and aspect as features. The literature on tense-aware MT presents cases where poor tense translation results in a failure to preserve temporal interpretation (e.g. Meyer (2014)), which we also observed. In contrast, we present here cases in which temporal interpretation is weakly preserved in spite of incorrect translation of tense and aspect.

- واشار الى ان الوزيرين "**سيقترحان**على نظيرهما الإسباني (خوسيه انطونيو الونسو) **عقد** اجتماع ثلاثي في القريب العاجل حول مكافحة تهريب المخدرات

- Ref: It also noted that both ministers "**will suggest** to their Spanish counterpart Jose Antonio Alonso **to hold** a tripartite meeting in **the near future** regarding combating drug trafficking."

- MT: He noted that the two ministers "**will suggest** on Spanish counterpart (Jose Antonio Alonso) tripartite meeting **was held** in the near future on the fight against drug trafficking."

In the reference translation we see that the meeting will take place after the suggesting, and during a period of time denoted by “the near future”. That Arabic similarly uses a past tense and infinite verb. It is fairly clear, however, that the past tense “was held” in the MT output is a tense error, as it seems likely that the meeting takes place during “the near future”. To the extent that one is familiar with Arabic to English translation, the MT system in question, and the training data, an incorrect tense translation (i.e. that “was held” should be “to hold”) is far more likely than a time expression denoting a past time being translated as “in the near future”. Another possibility is that a more accurate translation would put “in the near future” near the end of the sentence, so that it attaches to “drug trafficking” or “fight”. In this case we would have less evidence that “was held” has the wrong tense. It is not uncommon for an adverbial phrase to appear out of proper order. We also must face the problem that if tense translation errors are common, “will suggest” could have the wrong tense as well.

In spite of the various tense error possibilities, it seemed obvious that “was held” had the wrong tense, and so the source interpretation is weakly preserved. This is likely because in the context of the article, it does not make sense for the ministers to suggest that a meeting was held in the past. In another context, a similar construction might be more acceptable, as in:

- In his closing arguments, the defense attorney will suggest that a meeting was held for the purpose of framing his client in the near future.

Similarly, in the example below the lexical selection and semantics of the verb “to accuse” allow us to infer that an accusation of possession of weapons does not precede the possession itself, in spite of a tense translation error, resulting our categorizing this as a case of weak preservation by the MT. Note that replacing “to accuse” with “to convince” would not permit that inference, and we would not label this a case of weak preservation by the MT.

- والأدهى من ذلك وأمرّ سعيُ بعضهم إلى تلفيق أكاذيب **لاتّهام** هذا البلد أو ذاك **بامتلاك** أسلحة دمار شامل

- Ref: Even worse and more bitter is the attempt by some to fabricate lies, **accusing** this or that country of **possessing** weapons of mass destruction ...

- MT: Worse still, some of them and is seeking to fabricate lies **to accuse** this or that country **to possess** weapons of mass destruction ...

5 Conclusion: Proposal for New Resources & Shared Tasks

Given the exploratory analyses with the wide range of challenges in preserving temporal interpretation in MT of Arabic into English, we can now return to our earlier questions and conclude with a proposed research way forward:

(i) What resources are needed to assess the extent to which current MT systems preserve temporal information?

Parallel TimeBanks. There already exist corpora with temporal interpretation annotation in a variety of languages, [16] but we are only aware of one parallel effort, the Romanian TimeBank, built by translating English TimeBank and manually projecting the annotations to the Romanian translations. The creation of parallel TimeBanks, i.e., parallel corpora annotated for temporal interpretation, as in

[16] There exist TimeML annotated corpora in English, Spanish, French, Italian, Korean, Chinese, Indonesian, Brazilian Portuguese, Farsi, Estonian, Romanian.

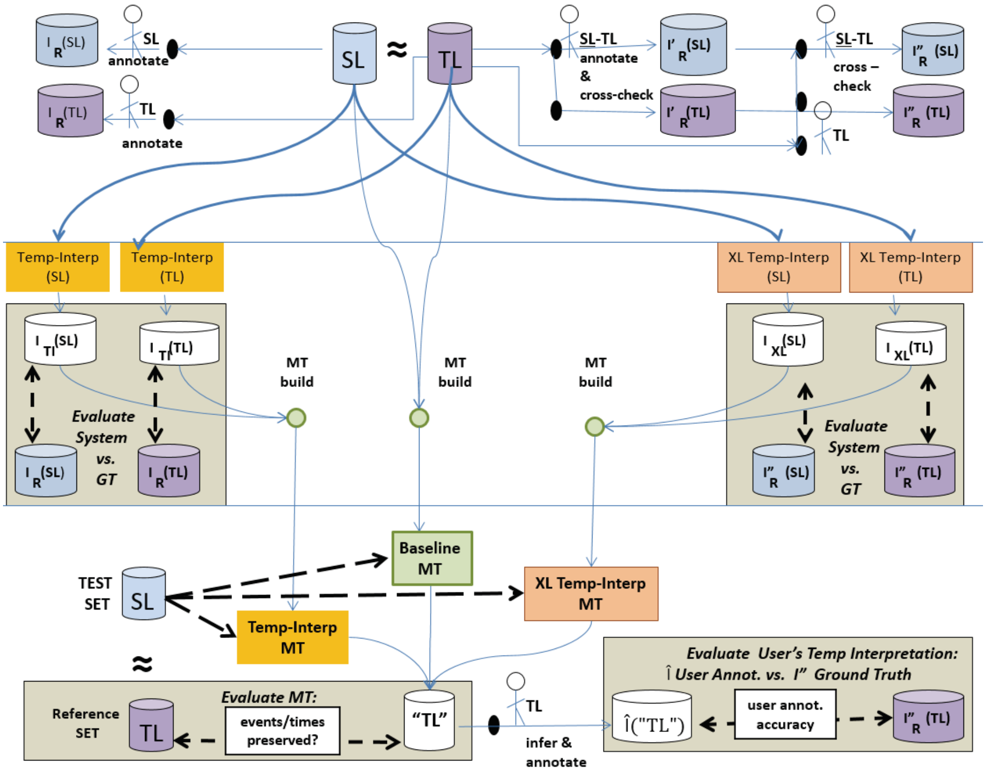

Figure 3: Workflow for Developing Parallel TimeBanks and Time-Aware Machine Translation.

figure 1a and the top of figure 3, will make it possible to support further development of temporal interpretation systems and new efforts toward time-aware MT, both shared tasks proposed below. We propose building a parallel English-Arabic TimeBank.

(ii) How might temporal interpretation extraction systems be integrated into the building of MT engines, and how will this impact the quality of MT output?

Shared Task: Temporal Interpretation Extraction Leveraging Parallel TimeBanks Shared tasks evaluating temporal interpretation algorithms have been conducted primarily in English. Each potential source language brings its own challenges due to the different ways in which temporal information is conveyed. We anticipate that the availability of parallel TimeBanks will encourage research on transfer learning for temporal interpretation extraction (see middle row of figure 3).

Shared Task: Time-Aware MT Leveraging Temporal Interpretation Extraction The desired results of the above proposed efforts are (1) a parallel English-Arabic TimeBank, and (2) publicly available tools for temporal interpretation extraction for Arabic text (those for English already exist). Given these resources, we propose a shared task to build MT systems that preserve the temporal interpretation of the source text. This task would address research questions pertaining to semantics-based MT as well as MT evaluation, both intrinsic and extrinsic (as in left and right evaluation boxes respectively in bottom row of figure 3).

References

Baran, E. (2013). *Chinese Verb Tense? Using English Parallel Data to Map Tense onto Chinese and Subsequent Tense Classification*. PhD thesis.

Bethard, S., Savova, G., Chen, W.-T., Derczynski, L., Pustejovsky, J., and Verhagen, M. (2016). Semeval-2016 task 12: Clinical tempeval. *Proceedings of SemEval*, pages 1052–1062.

Chang, A. X. and Manning, C. D. (2012). Sutime: A library for recognizing and normalizing time expressions. In *LREC*, pages 3735–3740.

Dorr, B. J. and Gaasterland, T. (2002). Constraints on the Generation of Tense , Aspect , and Connecting Words from Temporal Expressions. *Knowledge Creation Diffusion Utilization*, 1:1–47.

Forăscu, C. and Tufiş, D. (2012). Romanian timebank: An annotated parallel corpus for temporal information.

Ge, T., Ji, H., Chang, B., and Sui, Z. (2015). One Tense per Scene : Predicting Tense in Chinese Conversations. *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 668–673.

Gong, Z., Zhang, M., Tan, C. L., and Zhou, G. (2012). Classifier-based tense model for smt. In *COLING (Posters)*, pages 411–420. Citeseer.

Jones, D., Shen, W., Granoien, N., Herzog, M., and Weinstein, C. (2005). Measuring Translation Quality by Testing English Speakers with a New Defense Language Proficiency Test for Arabic. *International Conference on Intelligence Analysis*, page 6.

Katz, G. and Arosio, F. (2001). The annotation of temporal information in natural language sentences. *Proceedings of the workshop on Temporal and spatial information processing-Volume 13*, 13(1):15.

Klavans, J. L. and Chodorow, M. (1992). Degrees of Stativity: The Lexical Representation of Verb Aspect. *In Proceedings of {COLING} 1992*, pages 1127–1131.

Loaiciga, S., Meyer, T., and Popescu-Belis, A. (2014). English-french verb phrase alignment in europarl for tense translation modeling. In *LREC*, pages 674–681.

Matsuzaki, T., Fujita, A., Todo, N., and Arai, N. H. (2015). Evaluating Machine Translation Systems with Second Language Proficiency Tests. *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 145–149.

Meyer, T. (2014). Discourse-level features for statistical machine translation.

Meyer, T., Grisot, C., and Popescu-Belis, A. (2013). Detecting narrativity to improve english to french translation of simple past verbs. In *Proceedings of the 1st DiscoMT Workshop at ACL 2013 (51th Annual Meeting of the Association for Computational Linguistics)*, page 8.

Meyer, T., Hajlaoui, N., and Popescu-Belis, A. (2015). Disambiguating discourse connectives for statistical machine translation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 23(7):1184–1197.

Mirza, P. and Tonelli, S. (2014). Classifying Temporal Relations with Simple Features. *Aclweb.Org*, (2006):308–317.

Olsen, M., Traum, D., Van˜Ess-Dykema, C., and Weinberg, A. (2001). Implicit Cues for Explicit Generation: Using Telicity as a Cue for Tense Structure in a Chinese to English MT System. In *Proceedings of MT Summit VIII, Santiago de Compostella, Spain*.

Palmer, M., Gildea, D., and Kingsbury, P. (2005). The proposition bank: An annotated corpus of semantic roles. *Computational linguistics*, 31(1):71–106.

Pustejovsky, J., Castano, J. M., Ingria, R., Sauri, R., Gaizauskas, R. J., Setzer, A., Katz, G., and Radev, D. R. (2003a). Timeml: Robust specification of event and temporal expressions in text. *New directions in question answering*, 3:28–34.

Pustejovsky, J., Hanks, P., Sauri, R., See, A., Gaizauskas, R., Setzer, A., Radev, D., Sundheim, B., Day, D., Ferro, L., et al. (2003b). The timebank corpus. In *Corpus linguistics*, volume 2003, page 40.

Schaub, W. (1985). *Babungo*. Taylor & Francis.

Strötgen, J. and Gertz, M. (2010). Heideltime: High quality rule-based extraction and normalization of temporal expressions. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 321–324. Association for Computational Linguistics.

Styler IV, W. F., Bethard, S., Finan, S., Palmer, M., Pradhan, S., de Groen, P. C., Erickson, B., Miller, T., Lin, C., Savova, G., et al. (2014). Temporal annotation in the clinical domain. *Transactions of the Association for Computational Linguistics*, 2:143–154.

UzZaman, N., Llorens, H., Derczynski, L., Verhagen, M., Allen, J., and Pustejovsky, J. (2013). Semeval-2013 task 1: {T}empeval-3: Evaluating time expressions, events, and temporal relations. *Second joint conference on lexical and computational semantics (* SEM)*, 2(SemEval):1–9.

Verhagen, M., Gaizauskas, R., Schilder, F., Hepple, M., Moszkowicz, J., and Pustejovsky, J. (2009). The tempeval challenge: identifying temporal relations in text. *Language Resources and Evaluation*, 43(2):161–179.

Voss, C. R. and Tate, C. R. (2006). Task-based Evaluation of Machine Translation (MT) Engines : Measuring How Well People Extract Who , When , Where-Type Elements in MT Output. *Proceedings of the 11th Annual Conference of the European Associatio for Machine Translation (EAMT 2006)*, pages 203–212.

Ye, Y., Fossum, V. L., and Abney, S. (2006). Latent Features in Automatic Tense Translation between Chinese and English. (July):48–55.