Romanian TimeBank: An Annotated Parallel Corpus for Temporal Information

Corina Forăscu[1,2], Dan Tufiş[1]
[1] Research Institute in Artificial Intelligence, Romanian Academy, Bucharest
[1,2] Faculty of Computer Science, Alexandru Ioan Cuza University of Iaşi, Romania
16, General Berthelot, Iasi 700483, Romania
E-mail: corinfor@info.uaic.ro, tufis@racai.ro

Abstract

The paper describes the main steps for the construction, annotation and validation of the Romanian version of the TimeBank corpus. Starting from the English TimeBank corpus – the reference annotated corpus in the temporal domain, we have translated all the 183 English news texts into Romanian and mapped the English annotations onto Romanian, with a success rate of 96.53%. Based on ISO-Time - the emerging standard for representing temporal information, which includes many of the previous annotations schemes -, we have evaluated the automatic transfer onto Romanian and, and, when necessary, corrected the Romanian annotations so that in the end we obtained a 99.18% transfer rate for the TimeML annotations. In very few cases, due to language peculiarities, some original annotations could not be transferred. For the portability of the temporal annotation standard to Romanian, we suggested some additions for the ISO-Time standard, concerning especially the EVENT tag, based on linguistic evidence, the Romanian grammar, and also on the localisations of TimeML to other Romance languages. Future improvements to the Ro-TimeBank will take into consideration all temporal expressions, signals and events in texts, even those with a not very clear temporal anchoring.

Keywords: temporal information, annotation standards, Romanian language

1. Introduction

If during the 90s the temporal information only started to be brought to the attention of the NLP research community, nowadays the interest passed from the recognition and mark-up of temporal information in English texts (Mani et al., 2005) towards the standardisation of the temporal annotation (ISO-Time, 2009) in a multilingual context – Italian (Caselli, 2010), Korean, French (Bittar et al., 2011), German (Spreyer & Frank, 2008) – and the use of this information in almost all NLP areas: information extraction or information retrieval; question answering (dealing with questions like "when", "how often" or "how long", or temporally anchored questions as defined in QA competitions[1]); machine translation (translated and normalized temporal references; mappings between different behaviour of tenses from language to language; accurate translation memories); textual inference systems (to determine coreferential events); discourse processing: temporal structure of discourse and summarization (temporally ordered information, biographic summaries); medicine (summarizing data from temporal clinical databases, reasoning on temporal clinical data, monitoring intensive care patients, and planning and scheduling clinical routine activities).

The temporal elements in natural language are events – syntactically realized through sentences (mainly their syntactic head - the main verb), noun phrases, adjectives, predicative clauses or prepositional phrases, and temporal expressions – references to a calendar or clock system, expressed by noun, prepositional or adverbial phrases. These temporal elements can be found either explicit – in temporal expressions: *September 14, 2011*, *noon*, *one week*; events: *The reporter announced that the planned strike will start next Monday*. – or implicitly (*last week, next year, now, a few hours*) – in almost all acts of communication. These elements are linked so that the events can be positioned in time, either relatively with respect to other events or on an absolute time axis.

In order to have linguistic evidence and to study the temporal information in Romanian, we briefly present in section 2 our main steps towards porting the standard and creating a Romanian corpus: we used the TimeBank 1.2. corpus (Pustejovsky et al., 2006), together with the TimeML annotation scheme (Sauri et al., 2006); the translation, preprocessing and alignment of the corpus (Forăscu et al., 2007) are briefly presented in the same section. We automatically transferred the temporal annotation from English onto Romanian and evaluated this annotation import (Forăscu, 2008). Manual corrections and improvements were also used.

In section 3 we present some further improvements and additions (Forăscu, 2009; 2011) for the ISO-Time standard to be ported to Romanian, as well as Ro-TimeBank – the current version of the Romanian corpus.

The procedure used for the creation of the Romanian corpus is an appropriate one, given the success rate we obtained for the temporal transfer. The evaluation shows that this procedure can be easily used with other types of annotations or even with other language pairs. The paper shows, based on corpus-evidence, how well the temporal

[1] TAC (Text Analysis Conference, www.nist.gov/tac/), CLEF (Cross-Language Evaluation Forum, www.clef-initiative.eu//), TREC (Text REtrieval Conference, trec.nist.gov/)

theories can be applied to other languages, here with emphasis on Romanian. The corpus we created this way is publicly available through the META-SHARE[2] platform used in the METANET4U[3] project.

Future additions to the Ro-TimeBank corpus will consider also temporal elements not (yet) marked in the English version of the corpus. Most of these new elements, if they are not due to inevitable manual annotation mistakes, especially for the SIGNAL tag, have as rationale the fact that all sentences express an EVENT, through their main verb. New TIMEX3 tags can also be added to vague temporal elements (for example *not that long ago, once, begin of the week*).

2. Language Resources and the Creation of the Ro-TimeBank corpus

The existing Romanian LRs still do not support temporal annotation (Cristea & Forascu, 2006; Cristea, 2011) and the manual temporal annotation is very time consuming, expensive (Pustejovsky et al., 2002) and error-prone, including for Romanian (Forascu, 2011); therefore we decided to translate the English TimeBank and then to automatically import the original annotation from English into Romanian, based on the alignments between the parallel texts.

2.1 TimeML and TimeBank: the annotation standard and the English corpus

The TimeML mark-up language consists of a collection of tags intended to explicitly outline the information about the events reported in a given text (initially English texts, but currently with extensions to other languages), as well as about their temporal relations. The ISO-TimeML metadata standard marks:

- Events through the EVENT tag, to identify situations that happen or occur, states or circumstances in which something obtains or holds true. The MAKEINSTANCE tag, previously used (Sauri et al., 2006) for tracking the instances of a given event and for carrying the tense and aspect of the verb-denoted event, is no longer used in ISO-Time (2009).
- Temporal anchoring of events through the TIMEX3 tag (marking times – moments or periods of a day, dates, and durations: *Monday morning, two weeks, 9 a.m., noon, ...*), and the SIGNAL tag (function words indicating how temporal objects are to be related to each other: *during, at, twice, from, …*).
- Links between events and/or timexes through the temporal links (TLINK), aspectual (ALINK) and subordination (SLINK) links.

The TimeBank corpus consists of 183 news report documents, with XML markups for temporal information (TimeML 1.2. format), as well as other annotations. Even if the dimension of the corpus (4715 sentences with 10586 unique lexical units, from a total of 61042 lexical units) might be too small for robust statistical learning and the annotation might require corrections and improvements (Boguraev & Ando, 2006), the corpus is considered the reference corpus for temporal information.

2.2 Building the Romanian version of the TimeBank parallel corpus

The Romanian version of the TimeBank corpus was built following an *expand* procedure (Vossen, 1999): we translated the English corpus based on a minimal set of translation recommendations, designed also to enhance the alignment. The sentence alignment of the corpus was obtained as a direct output of the translation. In the 4715 sentences of the current version of the Romanian corpus there are 65375 lexical tokens, including punctuation marks, representing 12640 lexical types.

The English and Romanian raw texts were pre-processed in order to obtain the corpus in the format required by the lexical aligner. Using the TTL[4] module (Ion, 2007), the texts were tokenized, POS-tagged, lemmatized, and chunked. Then we used YAWA, a four stage lexical aligner based on bilingual translation lexicons and phrase boundaries detection to align words of a given bitext from Romanian to English (Tufis et al., 2005, 2006).

The automatic alignment performed on 181 files in the TimeBank parallel corpus produced 91714 alignments (25346 are NULL-alignments). Two files were not aligned because of a low translation quality.

We used the Romanian to English lexical alignment to transfer the XML markup from English to Romanian: we transferred into Romanian the TimeML mark-ups, as well as other mark-ups (for document format and structure information, sentence boundary information, and named entity recognition). The success rate for the import of the temporal mark-ups was 96.53%. The 3.47 % of non-transferred tags are due to missing translations (though the Romanian translation was a good and natural one), non-lexicalisations in Romanian, or missing alignments.

Using about 10% of the Romanian corpus, we performed a preliminary study (Forăscu, 2008) to analyze the situations of perfect transfer and compare them with some special situations (transfer with amendments or based on language specific phenomena, and impossible transfer). This study also laid the foundations for further improvements of the temporal annotations in Romanian, based on the last version of the TimeML standard, ISO-Time (2009).

[2] http://www.meta-share.eu/
[3] http://metanet4u.eu/
[4] http://ws.racai.ro/ttlws.wsdl

3. Ro-TimeBank – the Romanian corpus ISO-TimeML compliant

Following the TimeML development, we continued to adapt the Romanian corpus annotation to the ISO version of the standard and, meanwhile, we proceeded with the improvements (Forăscu, 2009) needed for the portability to Romanian of the ISO-Time standard (2009). We ground the Romanian specific rules and/or adaptations on the Romanian Academy grammar (GA, 2006). We also took into account the rules applied to other Romance languages: Italian (Caselli, 2010), French (Bittar et al., 2011). For all the tags in ISO-TimeML, we can apply almost the same rules from English. The main improvements concern the EVENT tag (Forăscu, 2011).

We opted to indicated whether an EVENT is a state (with the 'class' attribute having the value 'STATE'), instead of using the attribute 'type' to indicate if the EVENT is a state, a process or a transition. Our decision is compliant with TimeML simplified version, used in the AQUAINT and TempEval 1 and 2 corpora.

In order to reflect the Romanian tense system, with four tenses denoting the past, we propose to use two more values for the "tense" attribute of the EVENT tag, SIM_PAST for the "simple perfect" of the indicative (*perfect simplu* in Romanian) and PLUS_PAST for the „more than perfect" tense of the indicative (*mai mult ca perfect* in Romanian). For the „imperfect" tense (*imperfect* in Romanian), as well as for the „composed past" (*perfect compus* in Romanian) we use the value PAST; the distinction between these two tenses is realised through the value of the „aspect" attribute.

For the category of „aspect", we stick to the Romanian grammar and we include in the Romanian TimeML guidelines only the distinction between PERFECTIVE and IMPERFECTIVE verbs, manifested on the „imperfect" and „simple future" Romanian tenses on one side, and all the other tenses of the indicative mood, on the other side. For the EVENTS expressed by verbs in the present of indicative or by nouns, adjectives, prepositions or other part of speech, we use the value NONE for the „aspect" attribute.

Trying to keep compatibility between the ISO-Time standard (2009), the Romanian grammar (GA, 2006), as well as the other Romance ISO-TimeML standards (Italian, (Caselli, 2010) and French, (Bittar et al., 2011)), for the „mood" attribute of the EVENT tag we opted to include the values: CONDITIONAL/ IMPERATIVE/ SUBJUNCTIVE respectively for the conditional/ imperative/ subjunctive mood of the Romanian verbs. By default, the verbs in the indicative mood will have the NONE value for the „mood" attribute.

The "vform" attribute has four values in Romanian, corresponding to the non-personal moods, namely verbs in the INFINITIVE, GERUND, PARTICIPLE (the fourth value being the implicit NONE).

All the possibilities to assign values for the main attributes of the verb-denoting EVENTs are shown in Table 1.

mood	**tense**	**Romanian verb**	**tense attribute**	**mood attribute**	**vform attribute**	**aspect attribute**
Indicative	present	***vin***	PRESENT	NONE	NONE	NONE
Indicative	composed past	***am venit***	PAST	NONE	NONE	PERF
Indicative	simple perfect	***venii***	SIM_PAST	NONE	NONE	PERF
Indicative	more than perfect	***venisem***	PLUS_PAST	NONE	NONE	PERF
Indicative	imperfect	***veneam***	PAST	NONE	NONE	IMPERF
Indicative	future	***voi veni***	FUTURE	NONE	NONE	IMPERF
Indicative	future in the past	***voi fi venit***	FUTURE	NONE	NONE	PERF
Conditional	present	***aş veni***	PRESENT	CONDITIONAL	NONE	NONE
Conditional	perfect	***aş fi venit***	PAST	CONDITIONAL	NONE	NONE
Imperative		***vino***	PRESENT	IMPERATIVE	NONE	NONE
Subjunctive	present	***să vin***	PRESENT	SUBJONCTIVE	NONE	NONE
Subjunctive	perfect	***să fi venit***	PAST	SUBJONCTIVE	NONE	NONE
Infinitive		***a veni***	PRESENT	NONE	INFINITIVE	NONE
Participle		***venit***	PRESENT	NONE	PARTICIPLE	NONE
Gerund		***venind***	PRESENT	NONE	GERUND	NONE

Table 1: Values for verb-denoting events in Romanian

TimeML tags	# RO	# EN	% final
EVENTs	7926	7935	99.89
MAKEINSTANCEs	NA	7940	NA
TIMEXes	1414	1414	100.00
SIGNALs	669	688	97.24
TLINKs	6311	6418	98.33
SLINKs	2908	2932	99.18
ALINKs	262	265	98.87
TOTAL	**19490**	**19652**	**99.18**

Table 2: Transfer of TimeML mark-ups from English into the Romanian TimeBank corpus

Based on these considerations and on the ISO-Time standard, we developed the DTD for the Romanian ISO-Time, and, in a final processing step, we corrected the annotations in the Ro-TimeBank corpus in order to have the annotations compliant with the ISO version of the standard. At this stage we also included the two missing files in the Romanian corpus, obtaining the entire initial collection of 183 files.

Tables 2 and 3 show some basic statistics over the parallel corpus regarding the percent of the transferred (and/or corrected) tags, as well as statistics over the Ro-TimeBank corpus.

TLINKs

AFTER	890
BEFORE	1362
BEGINS	61
BEGUN_BY	69
DURING	290
ENDED_BY	174
ENDS	75
IAFTER	39
IBEFORE	34
IDENTITY	732
INCLUDES	569
IS_INCLUDED	1348
SIMULTANEOUS	668

SLINKs

CONDITIONAL	45
COUNTER_FACTIVE	47
EVIDENTIAL	1152
FACTIVE	395
MODAL	1263
NEG_EVIDENTIAL	6

TIMEXes

DATE	1164
DURATION	175
SET	12
TIME	63

ALINKs

CONTINUES	68
CULMINATES	30
INITIATES	93
REINITIATES	16
TERMINATES	55

EVENTs

ADJECTIVE	297
NA	27
NOUN	2350
OTHER	302
PREPOSITION	65
VERB	4885

Table 3: Statistics over the temporal tags in the Ro-TimeBank corpus

4. Conclusions

The research briefly presented in this paper proves that since manual annotating temporal information is very time-consuming and expensive, the automatic import, followed by a semi-automatic and manual evaluation and correction of the annotations represents a solution. This study shows how, based on corpus-evidence and localizations of the temporal annotation standard to other languages, temporal theories can be applied to languages other than the one they were initially created for, provided provided a parallel corpus exists and adequate processing tools are available.

The Ro-TimeBank corpus, finalized within the scope of the MetaNet4U project, is published on the Meta-Share [5] network of repositories of language data.

Future improvements to the Ro-TimeBank corpus will take into consideration all temporal expressions, signals and events in texts, even those with a not very clear temporal anchoring. The temporal elements not (yet) marked in the English TimeBank will be identified, over-passing the inevitable manual annotation mistakes, especially for the SIGNAL tag. For the EVENT tag we will consider that all sentences express an EVENT, through their main verb, even if the event might not have a very clear temporal anchor. New TIMEX3 tags will be added to mark also vague temporal elements.

5. Acknowledgements

This work has been supported by the MetaNet4U SPS European project under the grant no. #270893.

[5] http://www.meta-share.eu/

6. References

Bittar, A., Amsili, P., Denis, P., and Danlos, L. (2011). French TimeBank: An ISO-TimeML Annotated Reference Corpus. In *Proceedings of the 49th Annual Meeting of ACL*, Portland, Oregon, pp. 130--134.

Boguraev, B. and Ando, R. (2006). Analysis of TimeBank as a Resource for TimeML Parsing. *Proceedings of Language Resources and Evaluation - LREC-2006*, Genoa, Italy, pp. 71--76.

Caselli, T. (2010). *It-TimeML: TimeML Annotation Scheme for Italian* - Version 1.3.1. Technical Report, September 2010, ILC CNR Pisa, Italy.

Cristea, D. (2011). Romanian Linguistic Resources On Very Large Scale. In *Journal of Computer Science of Moldova*, Academy of Science of Moldova, vol. 19, nr. 2(56), ISSN 1561-4042, pp. 130--145.

Cristea, D. and Forăscu, C. (2006). Linguistic Resources and Technologies for Romanian Language. In *Journal of Computer Science of Moldova*, Academy of Science of Moldova, vol. 14, nr. 1(40), ISSN 1561-4042, pp. 34--73.

Forăscu, C., Ion, R., and Tufiş, D. (2007). Semi-automatic Annotation of the Romanian TimeBank 1.2. In *Proceedings of the RANLP 2007 Workshop on Computer-aided language processing - CALP*; Constantin Orăsan, Sandra Kuebler (Eds.). Borovets, Bulgaria, pp. 1--7.

Forăscu, C. (2008). GMT to +2 or How Can TimeML Be Used in Romanian. In *Proceedings of Language Resources and Evaluation - LREC-2008*, Marrakech, Morroco, May 26 - June 1, 2008, European Language Resources Association, ISBN 2-9517408-4-0.

Forăscu, C. (2009). A Romanian Corpus of Temporal Information - a Basis for Standardisation. In *Proceedings of Knowledge Engineering: Principles and Techniques Conference, KEPT 2009*, July 2-4 2009, Cluj-Napoca, Romania, ISSN 2067-1180, pp. 77--80.

Forăscu, C. (2011). Contributions to Romanian language processing through discourse analysis methods. (in Romanian). PhD thesis. Romanian Academy, Bucharest.

(GA, 2006) *The Grammar of the Romanian Language* (in Romanian *Gramatica limbii române. Vol. I Cuvântul, vol. II Enunţul*). Romanian Academy Publishing House.

Ion, R. (2007). Word Sense Disambiguation Methods Applied to English and Romanian. (in Romanian). PhD thesis. Romanian Academy, Bucharest.

(ISO-Time, 2009) *ISO: Language Resource Management – Semantic Annotation Framework (SemAF) - Part 1: Time and Events*. Secretariat KATS, 2009. ISO Report ISO/TC37/SC4 N269 (ISO/WD 24617-1).

Mani, I., Pustejovsky, J., and Gaizauskas, R. (eds.). (2005). *The Language of Time: A Reader*. Oxford University Press, ISBN-13: 978-0-19-926853-5.

Pustejovsky, J., Verhagen, M., Sauri, R., Littman, J., Gaizauskas, R., Katz, G., Mani, I., Knippen, B., and Setzer, A. (2006). *TimeBank 1.2*. Linguistic Data Consortium.

Pustejovsky, J., Belanger, L., Castaño, J., Gaizauskas, R., Hanks, P., Ingria, B., Katz, G., Radev, D., Rumshisky, A., Sanfilippo, A., Sauri, R., Setzer, A., Sundheim, B., Verhagen, M. (2002). NRRC Summer Workshop on Temporal and Event Recognition for QA Systems.

Sauri, R., Littman, J., Knippen, B., Gaizauskas, R., Setzer, A., and Pustejovsky, J. (2006). *TimeML Annotation Guidelines*, Version 1.2.1, January.

Spreyer K. and Anette Frank (2008). Projection-based Acquisition of a Temporal Labeller. In *Proceedings of the Third International Joint Conference on Natural Language Processing - IJCNLP 2008*, Hyderabad, India, pp. 489--496.

Tufiş, D., Ion, R., Ceauşu, A., and Ştefănescu, D. (2005). Combined Aligners. In *Proceedings of the ACL2005 Workshop on "Building and Using Parallel Corpora: Data-driven Machine Translation and Beyond"*, Ann Arbor, Michigan, pp. 107--110.

Tufiş, D., Ion, R., Ceauşu, A., and Ştefănescu, D. (2006). Improved Lexical Alignment by Combining Multiple Reified Alignments. In *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics - EACL 2006*, Trento, Italy, pp. 153--160.

Vossen, P. (1999). EuroWordNet. Building a Multilingual Database with Lexical-Semantic Networks for the European Languages. In *Proceedings of EUROLAN'99, 4th European Summer School on Human Language Technology.* Iasi, Romania. July 19-31, 1999.