A tale of temporal relations between clinical concepts and temporal expressions: towards a representation of the clinical patient's timeline

Cyril Grouin, MSc[1,2] Natalia Grabar, PhD[3] Thierry Hamon, PhD[4]
Sophie Rosset, PhD[1] Xavier Tannier, PhD[1,5] Pierre Zweigenbaum, PhD[1]

[1] LIMSI–CNRS, Orsay, France; [2] INSERM, UMR_S 872, Eq 20, Paris, France;
[3] STL CNRS UMR8163, Université Lille 1&3, Villeneuve d'Ascq, France; [4] LIM&BIO (EA3969), Université Paris 13, Bobigny, France; [5] Université Paris-Sud 11, Orsay, France

Abstract

This paper presents the experiments we made to process the temporal relations between clinical concepts and temporal expressions as part of our participation in the 2012 i2b2/VA challenge. In order to detect the clinical concepts, we reused and adapted the platforms we developed during the 2010 and 2011 i2b2/VA editions; those platforms integrate rules and machine-learning process. Moreover, we built models based upon Random Forest to identify the modality and the polarity of each concept. In order to identify the temporal expressions, we used the HeidelTime algorithm and made a few adaptations to deal with the specificities of the clinical documents. Finally, we split the set of temporal relations according to a series of distinct situations and built a series of models based upon decision trees with two strategies: the first one to give priority to precision, and the second one to balance recall and precision. After the official runs, we added more features, exhaustive enumeration of positive and negative instances at training time, and voting combination of five classifiers. On the first task (Event/Timex3 identification), our best submission achieved a 0.8307 global F-measure on the Event identification and a 0.8385 global F-measure on the Timex3. On the end-to-End task, we also achieved our best results on the third task, with a 0.4932 global F-measure on the Tlink identification; results on the Event/Timex3 are the same than the previous ones. Finally, on the Tlink task based upon the Event/Timex3 ground truth corpus, we achieved a 0.5471 global F-measure. The additional experiments performed after the official runs increased this F-measure to 0.5968.

Introduction

This year's i2b2/VA challenge focused on the temporal relations that occur between two elements being of two types: clinical concepts and temporal expressions. The aim of this challenge consists in identiying all events and temporal marks that would allow us to represent the clinical timeline of the patient. In 2010, the i2b2/VA challenge proposed to process the clinical concept identification issue;[1] three kinds of concepts were proposed: *problem, test, treatment*. In 2011, as part of the coreference resolution task, those concepts were completed with two new categories: *persons, pronouns*.[2] This year, the three original kinds were proposed in combination with three new categories: *clinical department, evidential, occurrence*; those six categories constitute the events we have to process.

In this paper, we describe the joint participation of two French teams that participated in the previous i2b2/VA challenges: the LIMSI team from 2009 to the current edition, and the Hamon & Grabar team from 2008 to 2010. In the current edition, we reused the systems we developed while participating in the previous editions to process the concept identification task: Caramba from the LIMSI team[3,4,5] and Ogmios from the Hamon & Grabar team.[6,7]

Corpus

The training corpus consists in 190 clinical records while the test corpus is composed of 120 files. Participants were asked to identify events, Timex3 and temporal relations with a few additional information: "type", "polarity" and "modality" for the events, "type" and "modifier" for the Timex3, and "type" for the temporal relations.

The annotation repartition within each value of related information is similar between both training and test corpora (see Table 1). Nevertheless, we noticed a huge unbalancy repartition between each related information: in the training corpus, 96.11% of the events are of modality *factual* (while the remaining events fall into one of the three others kinds and count for less than 3.89% of the total amount of events) and 92.97% of the events are associated with a *positive* polarity. The same proportion is observed in the test corpus. Similar unbalanced repartition are found in the Timex3 and Tlink related information, in both train and test corpora: Timex3 types are mainly of kind *date* without any available modifier.

Table 1: Annotation statistics in percentage on training and test corpora

	EVENT											
	Type						Polarity		Modality			
	Clinical Dept	Eviden-tial	Occu-rrence	Problem	Test	Treatment	Neg	Pos	Condi-tional	Factual	Possible	Proposed
TRAIN	6.05	4.49	19.95	**30.50**	15.76	**23.25**	7.03	**92.97**	0.87	**96.11**	1.71	1.31
TEST	5.39	4.38	18.38	**31.70**	15.99	**24.17**	6.84	**93.16**	0.99	**95.42**	2.23	1.36

	TIMEX3										TLINK		
	Type				Modifier						Type		
	Date	Duration	Freq	Time	Approx	End	Middle	More	NA	Start	After	Before	Overlap
TRAIN	**69.38**	17.19	10.52	2.91	10.98	2.41	0.08	0.30	**83.83**	2.41	9.56	**52.36**	**38.08**
TEST	**67.14**	18.74	10.82	3.30	9.88	1.71	0.11	0.28	**86.20**	1.82	9.84	**54.49**	**35.67**

Methods

Event processing

In order to identify the events, we reused the platforms we developed in the previous i2b2/VA challenges: Caramba and Ogmios. The adaptations we made mainly consisted in building new models for the machine-learning part of our systems in order to deal with the six categories of concepts proposed this year.

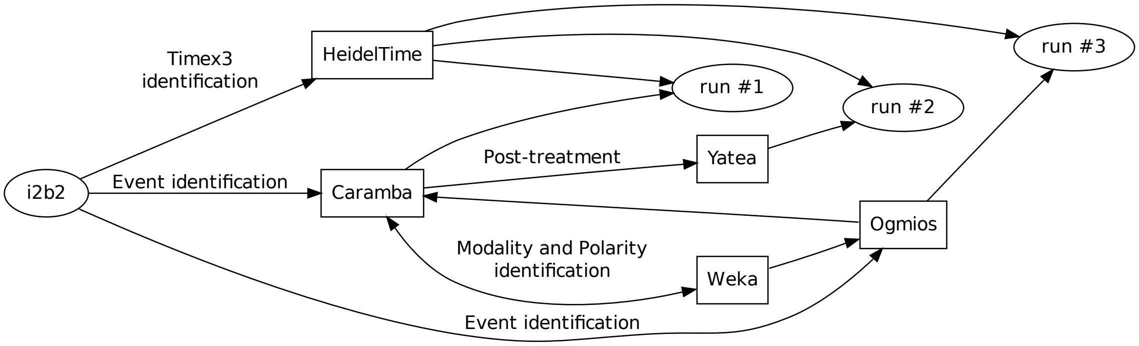

Figure 1: Configuration submissions in the Event/Timex3 and End-to-End tracks

Event identification

Caramba system. The Caramba system[3,4] relies on several tools that allow us to obtain various annotations we used as features to feed Wapiti,[8] a CRF classifier implementation. These features are shown in Table 2. In our experiments on the training corpus, we tested models generating up to 36 million features. Since CRFs may be prone to overfitting (e.g., compared to SVMs), we took care to include features that would lead to better generalization: syntactic tags and chunks, and semantic classes of various kinds. The output of the CRF was used as our first submission in both Event/Timex3 and End-to-End tracks (see Figure 1).

Table 2: Features for CRF-based event identification

- Section id among four sections we defined as follow: *admission date* (section #1), *discharge date* (#2), *history of present illness* (#3) and *hospital course* (#4);
- Morpho-syntactic tagging with the Tree Tagger (http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/)[9] and home-made noun phrase chunking based upon the previous tags;
- Morpho-syntactic tags projected from a specific lexicon of 62,263 adjectives and 320,013 nouns based on the UMLS Specialist Lexicon;
- Syntactic analysis with the Charniak McClosky biomedical parser (http://stanford.edu/~mcclosky/biomedical.html):[10] we used part-of-speech and chunk information derived from the parse trees;
- Semantic types and semantic groups from the UMLS (http://www.nlm.nih.gov/research/umls/);
- Semantic annotation (the six event types and other markers such as “anatomical part”, “localization”, “pre/post-examination”, “value unit”, etc.) with WMatch,[11] an analysis engine based upon regular expressions of words, rules and lexicons;
- Semantic annotations (the six event types and some specific semantic markers from the previous challenges such as “dosage”, “duration”, “mode of administration” (i2b2 2009), “pre/post-possible”, “pre/post-conditional”, “pre/post-problem”, “pre/post-treatment”, “pre/post-negation”, “pre/post-proposed”, etc. (assertion and concept markers from previous and this year challenges), provided by the Ogmios system;[12,7]
- Two series of unsupervised clusters obtained through Brown’s algorithm,[13] performed over the UMLS Metathesaurus[14] terms (multi-words expressions) and over the 2011 i2b2/VA Beth Israel and Partners Healthcare corpora. This corpus was selected because it was closest to the 2012 training corpus. It is probable that some test documents belong to this corpus, however since this processing is unsupervised and performed on the unannotated documents, it is good practice to do it (we should even have performed it again once we had the full set of unannotated test documents). Clustering was performed with code from Liang’s Master’s Thesis[15] (http://www.cs.berkeley.edu/~pliang/software/).

In order to refine the event extraction outputs of Caramba, we decided to use the term analysis performed by Yatea, a term extraction system.[16] For each multi-word concept found both by Yatea and Caramba, we added the boundaries of the concept specified by Yatea to the output of Caramba. This constitutes our second submission.

Ogmios system. Our third submission in Event/Timex3 and End-to-End tracks is the direct output of the Ogmios platform. Its configuration is similar to that defined for the 2010 i2b2/VA Challenge:[7]

1. POS tagging is performed with GeniaTagger,[17]
2. event identification relies on the TermTagger Perl module[*] and terminological resources,
3. a specific post-processing fitted to this year’s challenge selects and extends previously identified events.

We exploited three main terminological sources:

- 316,368 terms from the UMLS[14] which belong to several semantic axes related to:
  1. medical problems (B2.2.1.2.1 *Disease or Syndrome*, A2.2.2 *Sign or Symptom*, B2.3 *Injury and Poisoning*, A1.2.2 *Abnormality* and A1.1.5 *Bacteries*),
  2. tests (B1.3.1.1 *Diagnostic procedures* and B1.3.1.2 *Laboratory procedures*),
  3. treatments (B1.3.1.3 *Therapeutic or prevention procedures*,
  4. clinical departments (A2.7.1 *Health Care Related Organization*).
- 243,869 entries from RxNorm[18] used for the detection of medication names (treatments);

[*] http://search.cpan.org/~thhamon/Alvis-TermTagger/

- the annotations of the 2012 i2b2 training sets.

These resources have been manually checked and adapted in order to increase their coverage. Section titles have been also used for the categorization of events for instance, the *Admission Diagnosis* section usually contain medical problems.

The post-processing first performs the extension to the right of the event strings given that acceptable POS tags are accepted for this extension while the non-acceptable stopwords stop the extension. Then, it selects the correct event according to several indicators: (i) in case of concurrent annotations, the larger event string is preferred; (ii) event identified as a non i2b2 concept (*as well*, *M.D.*, etc.) is rejected; (iii) event occurring in section titles are removed except for the tests (*Serologies*) and occurrences (*Discharge Date*).

Event modality and polarity attributes

In order to process the polarity and modality on the events, we used the Weka toolbox (http://www.cs.waikato.ac.nz/ml/weka/). We chose to build two different models, one for the polarity and another one for the modality. We experimented several algorithms on the training corpus, using a 10-fold cross validation. Our models relie on the study of the 4 left and right tokens surrounding the identified concept. Out of our experiments, the *Random Forest* algorithm, based upon several decision trees, performed best: 0.994 on modality and 0.987 on polarity. Our models have been only trained on the outputs from Caramba. For the test, we used those models on the outputs from the event identification stage (either from Caramba or Ogmios systems) to process the modality and the polarity.

Timex3 identification

A preprocessing has been performed to identify the discharge date of a given document. This date has a central place because it is also considered as the date on which the document has been created. This information was helpful for Heideltime[19] to compute the reference time of relative expressions such as *two days later*. Moreover, some normalizations of the temporal expressions adopted by HeidelTime had also to be changed to fit into the i2b2 requirements: normalization of durations towards approximative numerical values rather than towards the undefined "X" value, external post-processing for some duration and frequency normalisations due to the limitation in the Heideltime arithmetic processor, etc.

Temporal relation identification

Official submission. We modeled the Tlink task as two distinct sub-tasks: first, deciding what pairs of elements deserve to be Tlinked, and then finding the appropriate relation between these elements. As summarized from the guidelines, the principle is to assign "*all Tlinks necessary to create a clinical timeline*". As annotating all pairs would be impractical and irrelevant, a system needs to reproduce this highly subjective process of deciding what relations participate in the elaboration of a comprehensive timeline. A consequence of this is that annotating correct relations between some elements may be judged as incorrect if the human annotator did not consider these relations as relevant.

Given two events or timexes that might be linked through a temporal relation, quite different situations arise depending on whether they are of the same kind (e.g., event-event) or not (event-timex), in the same sentence or paragraph or not, etc. We defined a total of 56 such situations according to the following dimensions:

- Sections *(admission date, discharge date, history of present illness, hospital course)*;
- Element types *(Timex3 or Event)*;
- Distance between elements *(same sentence, adjacent sentences in same section, more distant sentences)*;
- Number of Timexes between elements *(no Timex3 or at least one Timex3).*

Table 3: Features for decision-tree-based temporal link classification

- Text of the elements (if one of the 50 more frequent);
- Distance between the elements;
- Temporal or other prepositions between the elements;
- Number of other Timexes of Events between elements;
- All information coming from the Event or Timex3 annotations (modality, polarity, subtypes);
- For events, a subcategorization into *surgery*, *states*, *punctual events*, *locations*, *follow-up events*, or *others*, based on a lexical study of the development set.

Each combination of these dimensions defines a situation for which a distinct model is built. This is equivalent to defining an initial decision tree (with features based on the above four dimensions), after each leaf of which another classifier is applied. The intention was mainly to distinguish pairs of elements where assigning a Tlink was the rule, from those where it was the exception. Unfortunately, exceptions finally turned out to be the main rule. For situations where few positive examples existed in the training corpus, no model was used (and hence no Tlink was produced when applying the system).

Tlinks produced by one of these models might contradict previously predicted Tlinks. To address this issue, when applying each model, we sorted the predicted Tlinks in descending order of confidence. We then added each predicted Tlink in this order one at a time: for each one, its consistency with the current Tlink set was computed, and transitive closure on BEFORE and AFTER relations was computed. In case of inconsistency, the predicted Tlink was discarded.

We used the rule-based classifier J48, an implementation of the C4.5 decision tree learning algorithm,[20] as implemented in the Weka toolbox.[21,22] Classes were "BEFORE", "AFTER", "OVERLAP" and "NIL" (no relation). Features used for building all these models are shown in Table 3.

Two runs were submitted, giving preferences to different trade-offs between recall and precision. By applying all models with the same proportion of NILs as in the development set, we get a good precision but bad recall; on the other hand, imposing elements from the same sentence to be all Tlinked leads to well-balanced, both quite low, precision and recall.

Additional experiments. After the official test, we continued to investigate features, the handling of NIL relations, and different classifiers for the 56 models.

1. New features. For pairs of events located in the same sentence, syntactic dependencies may be relevant: we added those obtained with the Charniak McClosky parser, as converted into Stanford dependencies: a candidate pair of events or timexes is linked through a syntactic dependency if such a dependency links any word of the first to any word of the second.

The better performance of decision trees can be interpreted as indicating that combinations of features, rather than independent features, are meaningful predictors of Tlinks. Therefore we added several such combinations, beginning with the triples <source event type, syntactic dependency, target event type>, where source and target event types are the subtypes of events or timexes (i2b2/VA 2012 categories : the six event types, such as *evidential* and *occurrence*, and the four Timex3 types: *date*, *duration*, etc.).

2. Transitive closure and NIL relations. A possible drawback of the approach tested in our official submissions is that it applies classifiers to incomplete information: some positive instances of temporal relations can be inferred (through transitive closure) from those provided in the reference annotations of the training set, and should not be considered as negative examples. On the other hand, if we assume that annotators provided a virtually complete (i.e., complete after application of transitive closure) description of the temporal relations that can be found in the documents, all remaining pairs of events and timexes can be considered as negative instances. Although we know that this assumption is false to a certain extent, because such exhaustiveness was not required from annotators, we considered it would be worth testing how it fares as a principle for training classifiers.

We therefore tested the following setup. Transitive closure was run over all training Tlinks, using the Sputlink system included in the distributed i2b2/VA evaluation program, and all resulting Tlinks were used as positive instances. The cross-product of events and timexes of each document was used to generate a full set of candidate Tlinks, from which positive instances were filtered out to form the set of negative instances. On the contrary, transitive closure was not performed when applying the trained classifiers to the development or test data. The remaining part of the process was the same as that described above.

3. Different classifiers. The decision tree classifier obtained the best results globally with the new features and NIL relations too. However, it may be the case that different classifiers may be more or less adapted to different situations. We therefore evaluated the individual results of different classifiers (Naïve Bayes, decision trees, SVM, k nearest neighbors, logistic regression, random forest) for each of the 56 above-defined situations in the training set. We selected the best classifier for each situation and used the combination of the winners on the development and test data.

We also tested a different kind of combination of classifiers: using N different classifiers for each section and combining them through voting. We tested the combination of five classifiers: Naïve Bayes, decision trees, SVM, k nearest neighbors, and random forest, using average confidence as the combination operator. All classifiers were used through Weka.

Results and discussion

Track 1: Event/Timex3. On the event identification task, our best submission achieved a 0.8307 global F-measure using the Ogmios system (run #3, see Table 4). Moreover, we obtained a higher precision using Ogmios (0.7812) while Caramba allowed us to achieve a better recall (0.9576), probably due to our choice of features geared towards generalization.

The polarity and modality attributes processing have only been done on the outputs from Caramba. For the Ogmios event outputs, if the event was found by both systems, we used the modality and polarity attributes that were given for the Caramba's event. If an event was only identified by Ogmios, we gave this event the default modality and polarity values (i.e., *factual* and *positive*). In consequence, the modality and polarity scores are better on the runs #1 and #2 (Caramba) than the run #3 (Ogmios). Nevertheless, results are worse than the ones we should have obtained if we had given the default values on each event (0.9098<0.9316 on the polarity and 0.9142<0.9542 on the modality).

On the Timex3 identification sub-task, the HeidelTime algorithm performed well. It allowed us to achieve a 0.8385 global F-measure on the outputs from the Ogmios system. We noticed that some of our Timex3 outputs integrate erroneous normalization forms, such as "x" characters instead of a correct date. A complementary work on the normalization stage would be useful.

Table 4: Aggregated scores on the Event/Timex3 task

	Run	Precision	Recall	Average P&R	F-measure	Type	Polarity	Modality
EVENT	#1	0.6526	**0.9576**	0.7764	0.7762	**0.8607**	**0.9098**	**0.9142**
	#2	0.6339	0.9567	0.7627	0.7625	0.8598	0.9089	0.9139
	#3	**0.7812**	0.8869	0.8307	**0.8307**	0.7993	0.8401	0.8463
						Type	Val	Modifier
TIMEX3	#1	0.8605	0.8170	0.8382	0.8382	**0.7495**	**0.5363**	**0.7203**
	#2	0.8605	0.8170	0.8382	0.8382	0.7495	0.5363	0.7203
	#3	**0.8611**	0.8170	0.8385	**0.8385**	0.7478	0.5357	0.7203

Track 2: End-to-End. On the End-to-End task, our best submission is also the third one (based upon the Ogmios pipeline). We achieved a 0.4932 global F-measure on the temporal link subtask (see Table 5). On the second run

from this task, we noticed a malformed line in document 607.xml conducing the Python evaluation script to crash. A technical problem occurred while producing the final merge XML files, creating a broken line in the file. To perform the afore-mentionned evaluation, we deleted this line and closed the XML file.

The results on the Event/Timex3 identification subtask are the same than the previous ones (see Table 4) because this End-to-End global task was performed on our previous outputs.

Table 5: Aggregated scores on the End-to-End task (Tlink evaluation)

	Run	Precision	Recall	Average P&R	F-measure
TLINK	#1	0.5551	0.4114	0.4900	0.4726
	#2	0.5511	**0.4166**	0.4908	0.4745
	#3	**0.6155**	0.4115	**0.5175**	**0.4932**

Track 3: Tlink. On this last task, we only submitted two runs. We achieved our best results on the first submission with a 0.5471 global F-measure (see Table 6). While on the End-to-End task, our temporal links system was based upon our Event/Timex3 identification outputs, on this task, our system was applied on the Event/Timex3 ground truth. Depending on the used source, we achieved a better precision (0.6155) when processing our outputs (run #3 on Table 5) but a better recall (0.6015) when processing the ground truth corpus. This observation is compliant with the fact that on the ground truth corpus, each Event/Timex3 mention is supposed to be found, i.e., they are correct, which allows us to achieve a better recall for the Tlinks. For the first submission, we defined a strategy to obtain a good precision. Among the two submissions, we thus obtained our best precision (0.5105) on the first submission.

Table 6: Aggregated scores on the Tlink task (boldface shows the best results, italics show improvements over the official runs)

		Run	Precision	Recall	Average P&R	F-measure
TLINK	Official runs	#1	**0.5105**	0.5894	**0.5404**	**0.5471**
		#2	0.4940	**0.6015**	0.5341	0.5425
	Additional runs	*3 J48	*0.6436*	0.5321	*0.5977*	*0.5825*
		*4 SVM	**0.6587**	0.5110	*0.5955*	*0.5755*
		*5 RF	*0.5890*	0.5215	*0.5620*	*0.5532*
		*6 Vote	*0.6538*	0.5490	**0.6107**	**0.5968**

The best results of the additional tests that we performed after the official evaluation are shown as runs *3 through *6. They include the above-described new features, transitive closure and full cross-product of positive and NIL relations when training, and the use of either one classifier for all situations (*3–5), or one voting combination on five classifiers for all situations (*6). Our first experiments with different classifiers for different situations were not conclusive, more need to be performed in that line. The J48 decision tree (*3) was again the best single classifier in terms of F-measure ($F = 0.5825$); the SVM (*4, using LibSVM) obtained the best precision ($P = 0.6587$); the Random Forest meta-classifier (*5) also outperformed our official runs both in terms of precision and F-measure. Voting of these three classifiers plus k nearest neighbors (with $k = 1$) and Naïve Bayes outperformed all other setups in terms of average P&R (0.6107) and F-measure ($F = 0.5968$), increasing by 5 points our best official run: 3.5 points through the new scheme and features, and an additional 1.4 points through voting combination. These improvements were obtained by boosting precision (+14 points) at the price of a reduced recall (–4 or 5 points).

We plan to perform experiments with more syntactic features and more combined features to try and push these improvements further.

Conclusion

In this paper, we presented the experiments we made to process the temporal relations between clinical concepts and temporal expressions as part of our participation in the 2012 i2b2/VA challenge.

We reused and adapted the *Caramba* and *Ogmios* platforms we developed during the 2010 and 2011 i2b2/VA editions to detect the clinical concepts. Those platforms integrate both rules and machine-learning process. In order to identify the modality and the polarity of each concept, we built models based upon the Random Forest meta-classifier. The identification of the temporal expressions has been made using the HeidelTime algorithm. We made a few adaptations to deal with the specificities of the clinical documents.

Finally, to process the temporal relations, we built specific models for a series of situations, based upon decision trees, with a post-filter to remove inconsistent relations. Our two official runs applied two strategies: the first one to give priority to precision, and the second one to balance recall and precision. In additional experiments, we increased the F-measure by 5 points through additional syntax-based and combined features, exhaustive enumeration of positive and negative instances at training time, and voting combination of five classifiers.

On the first task (Event/Timex3 identification), our best submission achieved a 0.8307 global F-measure on the Event identification and a 0.8385 global F-measure on the Timex3. On the end-to-End task, we also achieved our best results on the third run, with a 0.4932 global F-measure on the Tlink identification; results on the Event/Timex3 are the same as the previous ones. Finally, on the Tlink task based upon the Event/Timex3 ground truth corpus, we achieved a 0.5471 global F-measure in official runs, and increased it to 0.5968 in posterior experiments.

Acknowledgement

This work has been partly done within the framework of the Quaero project funded by Oseo, French State Agency for Research and Innovation, and by the Accordys project,[†] funded by ANR under grant number ANR-12-CORD-0007-03.

We thank the organizers for the hard work they did!

References

1. Ozlem Uzuner, Brett R. South, S. Shen, and Scott L. Duvall. 2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text. *J Am Med Inform Assoc*, 18(5):552–556, 2011.

2. Ozlem Uzuner, John Pestian, and Brett South. The i2b2/VA 2011 challenge. In *i2b2 Workshop Proc*, 2011.

3. Cyril Grouin, Asma Ben Abacha, Delphine Bernhard, Bruno Cartoni, Louise Deléger, Brigitte Grau, Anne-Laure Ligozat, Anne-Lyse Minard, Sophie Rosset, and Pierre Zweigenbaum. CARAMBA: Concept, assertion, and relation annotation using machine-learning based approaches. In *i2b2/VA Workshop Proc*, 2010.

4. Anne-Lyse Minard, Anne-Laure Ligozat, Asma Ben Abacha, Delphine Bernhard, Bruno Cartoni, Louise Deléger, Brigitte Grau, Sophie Rosset, Pierre Zweigenbaum, and Cyril Grouin. Hybrid methods for improving information access in clinical documents: Concept, assertion, and relation identification. *J Am Med Inform Assoc*, 18(5):588–593, 2011.

5. Cyril Grouin, Marco Dinarelli, Sophie Rosset, Guillaume Wisniewski, and Pierre Zweigenbaum. Coreference resolution in clinical reports. The LIMSI participation in the i2b2/VA 2011 challenge. In *i2b2/VA Workshop Proc*, 2011.

6. Thierry Hamon and Natalia Grabar. Linguistic approach for identification of medication names and related information in clinical narratives. *J Am Med Inform Assoc*, 17(5):549–54, 2010.

[†] Accordys: *Agrégation de Contenus et de COnnaissances pour Raisonner à partir de cas de DYSmorphologie fœtale*, Contents and Knowledge Aggregation for Case-based Reasoning in the field of Foetal Dysmorphology (ANR 2012-2015).

7. Thierry Hamon, Amandine Périnet, Jérôme Nobécourt, and Natalia Grabar. Linguistic and semantic annotation for information extraction and characterization. In *i2b2/VA Workshop Proc*, 2010.

8. Thomas Lavergne, Olivier Cappé, and François Yvon. Practical very large scale CRFs. In *ACL Proc*, pages 504–513, 2010.

9. Helmut Schmid. Probabilistic part-of-speech tagging using decision trees. In *Proc of International Conference on New Methods in Language*, 1994.

10. David McClosky, Eugene Charniak, and Mark Johnson. Automatic domain adapatation for parsing. In *NAACL–HLT Proc*, 2010.

11. Olivier Galibert. *Approches et méthodologies pour la réponse automatique à des questions adaptées à un cadre interactif en domaine ouvert*. PhD thesis, Université Paris-Sud, Orsay, 2009.

12. Thierry Hamon and Natalia Grabar. Concurrent linguistic annotations for identifying medication names and the related information in discharge summaries. In *i2b2/VA Workshop Proc*, 2009.

13. Peter F. Brown, Vincent J. Della Pietra, Peter V. de Souza, Jennifer C. Lai, and Robert L. Mercer. Class-based n-gram models of natural language. *Computational Linguistics*, 18(4):467–479, 1992.

14. Olivier Bodenreider. The unified medical language system (UMLS): integrating biomedical terminology. *Nucleic Acids Research*, 32:267–270, 2004.

15. Percy Liang. Semi-supervised learning for natural language. Master's thesis, MIT, 2005.

16. Sophie Aubin and Thierry Hamon. Improving term extraction with terminological resources. In *Advances in NLP (5th International Conference on NLP, FinTAL)*, 2006.

17. Yoshimasa Tsuruoka, Yuka Tateishi, Jin-Dong Kim, Tomoko Ohta, John McNaught, Sophia Ananiadou, and Jun'ichi Tsujii. Developing a robust part-of-speech tagger for biomedical text. In *Proceedings of Advances in Informatics - 10th Panhellenic Conference on Informatics*, LNCS 3746, pages 382–392, 2005.

18. National Library of Medicine, Bethesda, Maryland. *RxNorm, a standardized nomenclature for clinical drugs*, 2009. Available at www.nlm.nih.gov/research/umls/rxnorm/docs/index.html.

19. Jannik Strötgen and Michael Gertz. Temporal tagging on different domains: Challenges, strategies, and gold standards. In *LREC Proc*, 2012.

20. R. Quinlan. *C4.5: Programs for Machine Learning*. Morgan Kaufman Publishers, 1993.

21. Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. The WEKA data mining software: An update. *SIGKDD Explorations*, 11(1), 2009.

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