Disease and Disorder Template Filling Using Rule-Based and Statistical Approaches

Thierry Hamon[1,2], Cyril Grouin[1], and Pierre Zweigenbaum[1]

[1] LIMSI–CNRS, Campus universitaire d'Orsay, bât. 508, rue John von Neumann, F-91405 Orsay, France [2] Université Paris 13, Villetaneuse, France

Abstract. We present the participation of LIMSI in Task 2 of the 2014 ShARe/CLEF eHealth Evaluation Lab. We used a hybrid approach based on a rule-based system and supervised classifiers depending on the properties of the attributes. The rule-based system identified course, severity and body location attributes based on the annotations of the training set and resources obtained from the UMLS. The Heideltime system was used to identify the dates. A MaxEnt model was trained to detect negation and uncertainty based on the disorder and surrounding words. A Decision Tree detected the relation to document time based on the position of the disorder in the document and on the words in the current sentence. Our system obtained a global 5th position out of ten ranked teams (accuracy of 0.804), and ranked 2nd for the detection of the relation to document time (accuracy of 0.322).

Keywords: natural language processing, medical records, machine learning

1 Introduction

Medical records contain a wealth of information on patients covering their hospital stays, including health conditions, diagnoses, performed tests, treatments. A large part of this information is held in free text. Information extraction from free text medical records now has a long history [5, 16, 22]. While these earlier text analysis systems aimed at a detailed representation of text contents, more recent shared tasks (e.g., i2b2/VA 2010 [21]) have generally handled medical entities such as medical problems (aka disorders) as atomic. This was the case of the 2013 ShARe/CLEF eHealth T2 task [19] which required to detect disorders spans and their concept unique identifiers (CUIs).

In contrast, the 2014 ShARe/CLEF eHealth T2 shared task [11] focuses on the attributes of such disorders. According to the task guidelines [4], the attributes can be divided into five categories: anatomical information concerning the location of the disorders in the body (BL), assertions on the disorder concerning negation and uncertainty indications (NI, UI), clinical information describing the disorder severity (SV) and its course (CC), contextual information to identify the subject who experiences the disorder (SC) and the condition (CC) in

which the disorder exists, and temporal information including the time expression related to the disorder (TE) and the temporal relation between the disorder and the time of the document (DT).

General methods to perform this task include knowledge-based methods which specify in which condition an attribute should be recognized for a given disorder mention, e.g. by detecting terms in lexicons or by matching lexico-syntactic patterns; and machine-learning based methods which learn to detect the presence of an attribute from a feature representation of each disorder mention. The overall approach of the LIMSI team is hybrid: depending on the properties of the attributes, we used rule-based methods relying on linguistic and terminological resources (BL, SV, CC, TE) or supervised classifiers (NI, UI, DT) to identify and normalise disorder attributes.

This paper is organized as follows. In Sec. 2 we present related work on attribute recognition in clinical texts. Then, we detail the materials and methods we used according to the attributes in Sec. 3. Results are presented and discussed in Sec. 4.

2 Related Work

Negation (NI), uncertainty (UI), subject (SC) and conditional existence (CO) were part of the task to address in the i2b2/VA 2010 [21] and i2b2 2012 [18] challenges, under one category called "assertion". Most of the top-ten ranked systems obtained a high F-measure around 0.93, using supervised methods or hybrid systems.

Temporal expressions (TE) and relations (including that in DT) were addressed in the i2b2 2012 challenge [18], albeit in a slightly different way. Temporal expressions were to be detected anywhere in the text and did not need to be related to a specific disorder. Each event needed to be anchored to the patient timeline through a temporal relations. Events included "problems", which were close to the "disorders" addressed in the present task. Besides, temporal relations targets could to be any event or temporal expression, including the admission or discharge dates. The top-ten systems obtained F-measures of 0.45–0.66 for finding the normalized value of a temporal expression (Timex3), and 0.43–0.69 for temporal relations between any pair of events.

Anatomical parts (BL) have been included in manual annotations in a few corpora, including MiPACQ [1] and Quaero [13]. Roberts *et al.* [14] report an F-measure of 0.86 for extracting the anatomical site of an actionable finding in radiology reports.

3 Materials and Methods

3.1 Data

The corpus used for the 2014 ShARe/CLEF eHEALTH evaluation lab consists of de-identified plain text EMRs from the MIMIC II database, version 2.5 [15].

Table 1. Description of the corpora

	Training	Test
Documents	299	133
Words	182,056	153,558
Distinct CUIs	1,356	1,141

Table 2. Statistics of classes for each attribute from the training corpus

Attribute	Classes
Body Location (BL)	null (5,131), C0000726 (403), C0003501 (295), C0225897 (271), C0278454 (215), C0026264 (208), C0018787 (183), C0817096 (172), C0024109 (134), C0031050 (127), C0007226 (103), C0018792 (101), etc.
Conditional Class (CO)	false (10,993), true (560), null (1)
Course Class (CC)	unmarked (10,887), increased (234), decreased (186), improved (101), worsened (67), resolved (63), changed (12), null (3), no (1)
DocTime Class (DT)	overlap (6,851), before_overlaps (2,814), before (1,391), after (442), unknown (55)
Generic Class (GC)	false (11,553), null (1)
Negation Indicator (NI)	no (9,349), yes (2,205)
Severity Class (SV)	unmarked (10,344), moderate (671), severe (410), slight (128), null (1)
Subject Class (SC)	patient (11,467), family_member (72), other (13), donor_other (1)
Temporal Expression (TE)	none (8,094), date (3,266), duration (131), time (62), overlap (1)
Uncertainty Indicator (UI)	no (13,539), yes (1,014), null (1)

The EMR documents were extracted from the intensive-care unit setting and included discharge summaries, electrocardiography reports, echography reports, and radiology reports.

The training set contained 299 documents and a total of 182,056 words, while the test set contained 133 documents and a total of 153,558 words (see Tab. 1). In Tab. 2, we give a few statistics for each attribute in the training corpus.

3.2 System Design

Three types of methods were used in our system depending on the properties of an attribute:

– (*i*) because unbalanced distributions are hard to process (see Tab. 2), attributes with a very large majority class (SC, CO, GC) were not addressed minimally: the majority value was systematically returned for such attributes;[1]

[1] Subject Class=*patient*; Conditional Class=*false*; Generic Class=*false*.

– Attributes with more variation were handled with either
  • (*ii*) human-designed resources and rules if clear clues could be collected and organized to make a decision for such attributes (CC, BL, SV, TE);
  • or (*iii*) supervised classification if some of the clues played a less categorical role in decision-making (NI, UI, DT).

We detail below the methods used for attributes with more variation: rule-based detection of temporal expression (Sec. 3.3), resource-based detection of body location, severity and course (Sec. 3.4), supervised detection of negation and uncertainty (Sec. 3.5), and supervised detection of DocTime class (Sec. 3.6).

3.3 Rule-Based Detection of Temporal Expression

To identify the temporal expressions, we used the rule-based temporal tagger Heideltime [17] that we tuned for clinical texts during the 2012 i2b2 challenge [9]. This tuned version of Heideltime includes linguistic patterns specific to medical and especially clinical temporal expressions, such as *postoperative day four*, *day of life*, etc. For the CLEF-eHealth challenge, we only used the date expressions Heideltime recognises, since the other temporal expressions (duration and time) were too rare in the training set and their recognition decreased the performance of our system on that corpus.

3.4 Resource-Based Detection of Body Location, Severity and Course

The recognition of the terms for the attributes body location (BL), course (CC), and severity (SV), was based on resources specifically built for each attribute.

Since course and severity were marked with fairly regular clues in the training corpus (see Tab. 2), we used the annotations of the training set as resources to identify linguistic expressions related to these two attributes.

Terminological resources used for the recognition of terms referring to body locations were built from the training annotations as well as from UMLS Metathesaurus terms from selected source vocabularies. During preliminary experiments on the training set, we observed that terms found in some UMLS vocabularies tend to decrease the quality of the annotation. For this reason, we only considered UMLS terms obtained from four source vocabularies:

– Health Level Seven Vocabulary (HL7),
– Metathesaurus Forms of FDA National Drug Code Directory (FDA),
– University of Washington Digital Anatomist (UWDA),
– and UMLS Metathesaurus specific terms (MTH).

Ambiguous annotations such as “a” or “his” occurring in the training annotations were removed from the body location resource we used. We considered the CUIs as fine-grained semantic tags associated to the BL terms.

These resources were used by the TermTagger Perl module[2] to recognise SV and CC mentions and BL terms. The clinical texts were also semantically tagged with the CUIs associated to BL terms. Term tagging is integrated in the Ogmios platform [10] which first performs POS-tagging with GeniaTagger [20]. For each disorder mention, a post-processing step selected the BL, SV and CC terms found in the sentence where the disorder occurs.

3.5 Supervised Detection of Negation and Uncertainty

System Description. Based upon an empirical analysis of the training corpus, we prepared a list of clues we found relevant to detect negation and uncertainty:

- Negation clues: *negative, no, not, without, denies, deny*;
- Uncertainty clues: *appear, assess, could, evaluate, likely, may, possible, possibility, possibilities, prior, probable, questionable, somewhat, suggesting, suspicion, unknown*. We also marked the *PATIENT/TEST* subsection header as an uncertainty clue, since we observed that disorders in this subsection were associated with an uncertainty indicator in the training corpus.

They were then used to mark as negated or uncertain the part of a sentence following such a clue, thereby implementing a simplified scope detection method.

In order to detect negated and uncertain disorders, we designed two distinct models based upon the Maximum Entropy framework [2, 7] as implemented in the Wapiti toolkit[3] [12]: one model for negation identification (NI), and one for uncertainty identification (UI). Our models rely on both surface and external features:

- Surface features: (*i*) the whole entity, (*ii*) each token from the entity as a bag of words, (*iii*) the capitalization of each token among four schemas (all in upper case, all in lower case, combination of upper and lower case, not relevant), and (*iv*) the three tokens preceding the entity to process;
- External features: (*i*) the Concept Unique Identifier (CUI) of the whole entity as found in the UMLS Metathesaurus [3], and (*ii*) whether the part of the sentence where the entity is found is negated or uncertain, based upon negation and uncertainty clues found before the current entity.

Example. For the entity *"Allergies to Drugs"* in the sentence *"Patient recorded as having No Known Allergies to Drugs"*, we used the following features:

- Whole entity: *Allergies_to_Drugs*;
- Tokens from the entity (bag of words): *Allergies, to, Drugs*;
- Capitalization of each token: *Mm, mm, Mm* (i.e., the first and third tokens combine lower and upper case while the second token is only in lower case);
- Three tokens preceding the entity (bag of words): *having, No, Known*;
- CUI of the entity: *C0013182*;
- Part of the sentence where the entity is found being marked as negated or uncertained: *NEG* (the clue "no" was found in the left context of the entity).

[2] `http://search.cpan.org/~thhamon/Alvis-TermTagger/`
[3] `http://wapiti.limsi.fr/`

3.6 Supervised Detection of Temporal Relation to Document Time

The Document Time attribute encodes the temporal relation between a disorder and the date of the document. Clinical reports often follow the chronological order of reported events. A study of the training corpus confirmed this principle. It also showed that the document structuring into sections often goes together with specific distributions of temporal relations in each section. For example, typically, the Chief Complaint section covers past disorders, the Pertinent Results section describes disorders which overlap the hospital stay, and the Medications on Discharge section mention disorders that may occur after discharge. We therefore emphasized the use of document structure as an important clue to determine the temporal relation of a disorder. To do so, we compiled a list of the most frequent section headers found in the training corpus, and encoded it as patterns to detect 26 section types. We also modeled the position of a disorder in a document as both its character offset and its relative position by cutting the text into five equal-sized bins. In principle, verb tense should also contribute to relative time positioning; unfortunately we could not test it for want of time.

We addressed this sub-task as a supervised classification task with four classes: BEFORE, BEFORE_OVERLAPS, OVERLAP, AFTER. For each disorder, we collected the following features:

- position in the text (absolute and discretized in five equal bins);
- document type, section type, and their conjunction;
- tokens in the sentence, as a bag of words.

The conversion of sentences into bags of words considered the absence or presence of each word with at least 10 occurrences in the set of sentences for each class.

We tested several classifiers of the Weka toolkit [8] by training and testing them in ten-fold cross-validation on the training set (see Tab. 3): majority class (ZeroR: OVERLAP), set of rules operating on only one feature (OneR: operates on conjoined feature document_type+section_type), Naïve Bayes (NB), Decision Tree (J48, confidence threshold 0.4, minimal number of instances per leaf 10), k Nearest Neighbors (kNN with $k = 1, 3, 5$), SVM (SMO with polynomial kernel). The best results on the training set before the submission were obtained by the

Table 3. Document Time attribute: performance on training set with various classifiers (ten-fold cross validation). Results marked in bold have been obtained after the submission.

Classifier	ZeroR	OneR	NB	J48	kNN: $k = 1, 3, 5$	SVM
Accuracy	0.593	0.739	0.788	0.814	**0.842, 0.823, 0.813**	**0.844**

decision tree, which was therefore used as the classifier for the test corpus. We can see in Tab. 3 that although slightly better results could be achieved after the submission with similarity-based classifiers such as kNN or SVM, the obtained range seems to be close to the maximum that can be obtained with the current features.

3.7 Submissions

We submitted two system outputs based upon the predictions performed by the previous systems. The only difference between the two submissions pertained to the Temporal Expression attribute: the first submission only focused on classes *date* and *none* which were most often found with this attribute (see Tab. 2), while the second submission also took into account the less represented *time* and *duration* classes.

4 Results and Discussion

4.1 Evaluation Metrics

The official evaluation measure is the overall average accuracy, where the accuracy of each attribute is defined as

$$Accuracy = \frac{Correct}{Total} \tag{1}$$

where *Correct* is the number of entities with correctly predicted value and *Total* is the number of entities in the gold standard annotations.

4.2 Results on the Training Set

To estimate the performance of the system, two methods can be used. Admittedly, the method which best helps predict future results on unseen data consists in using cross-validation, i.e., preparing a system based on a subset of the training data and testing it on the rest, repeating the process on different splits of the training data. This is easy to do for machine-learning systems: Table 3 showed the accuracy obtained on the Document Time attribute with ten-fold cross-validation on the training set.

For knowledge-based systems however, it is more cumbersome to use multiple splits of the same dataset since the human knowledge engineer / system developer cannot "forget" the data she has seen in a previous split to prepare a new version of the system. Working on one split is possible although less predictive of future results. What we present here is simply the application of the system modules prepared on the training set and tested on the training set itself. While this is not in principle highly predictive of future results, it often gives an idea of where the system stands. Table 4 shows the overall results obtained this way, while Table 5 provides detailed information for each attribute. The obtained results are likely to be optimistic, especially for machine-learning systems, which generally tend to overfit the training data. We return to them when examining the results on the test data.

Table 4. Results on the training set

Submission	Accuracy	F-measure	Recall	Precision
#1	0.884	0.684	0.674	0.693
#2	0.882	0.682	0.677	0.686

Table 5. Detailed results on the training set for each attribute. The δ value for each attribute represents its difference to the best system.

Method	Attribute	Accuracy	F-measure	Recall	Precision
Default value	GC	1.000	0.000	0.000	0.000
	SC	0.992	0	0.000	0.000
	CO	0.950	0	0.000	0.000
Resource-based	SV	0.877	0.420	0.573	0.332
	CC	0.859	0.388	0.832	0.253
	BL	0.511	0.375	0.404	0.350
Rule-based	TE (#1)	0.692	0.071	0.040	0.289
	TE (#2)	0.678	0.104	0.065	0.258
MaxEnt	NI	0.966	0.905	0.827	0.998
	UI	0.989	0.936	0.884	0.994
Decision tree	DT	N/A	N/A	N/A	N/A

4.3 Global Results

Table 6 shows the official results we achieved on the test set. Our first submission ranked 6th out of 12 submissions, and 5th out of 10 participants. We can see that the overall accuracy is not much lower than that obtained on the training data (-0.08)—however, recall, precision, and F-measure are much lower (they are divided by two).

Table 6. Official results on the test set

Submission	Accuracy	F-measure	Recall	Precision
#1	0.804	0.315	0.303	0.330
#2	0.801	0.315	0.290	0.333

4.4 Detailed Results per Attribute

Table 7 displays the detailed results we achieved on the test set for each attribute. For the Temporal Expression (TE) attribute, we indicate the results we achieved for both submissions. For the other attributes, there is no difference between submissions #1 and #2. We also indicate the δ value between our submissions and the best submission for each attribute.

Table 7. Detailed results on the test set for each attribute. The δ value for each attribute represents its difference to the best system.

Method	Attribute	Accuracy	δ	F-measure	Recall	Precision
Default value	GC	1.000	-0.000	0.000	0.000	0.000
	SC	0.984	-0.011	0.000	0.000	0.000
	CO	0.936	-0.042	0.000	0.000	0.000
Resource-based	SV	0.900	-0.082	0.395	0.282	0.663
	CC	0.853	-0.118	0.281	0.172	0.765
	BL	0.504	-0.293	0.277	0.248	0.313
Rule-based	TE (#1)	0.839	-0.025	0.092	0.186	0.061
	TE (#2)	0.806	-0.058	0.126	0.156	0.106
MaxEnt	NI	0.902	-0.067	0.722	0.879	0.612
	UI	0.801	-0.159	0.026	0.018	0.044
Decision tree	DT	0.322	-0.006	0.322	0.322	0.322

The results for attributes handled through default values or resources are very close to those obtained on the training set. Surprisingly, the accuracy obtained for the rule-based TE attribute is much better on the test set than on the training set. A possible explanation could be related to the fact that the test set only contained discharge summaries whereas the training set also contained echography, ECG and radiography examination reports, with maybe more regular temporal expressions in the discharge summaries.

4.5 Discussion

For highly unbalanced attributes (GC, SC, CO), the decision not to process these attributes and to select the majority class instead proved good: we achieved our better accuracy values on these three attributes. We notice that most teams did the same for GC (which did not vary at all in the training set), four other teams did the same for CO (ranking #5 before 2 teams), and one other team did so for SC (ranking #5 before 4 teams). For SC, the distance to the best team, which obtained near-perfect results, is only 0.009; for CO, it was 0.042: there is more to gain there with a more precise strategy.

For attributes relying on lists, the resource-based approach obtained moderate results. For example, the CUIs for the Body Location attribute encompass a high number of distinct values, which makes it difficult to detect with high accuracy. The simple dictionary-based method that we used to detect BL mentions with a co-occurrence based method to associate them to a disorder underperformed compared to other participants (-0.29 wrt. the best system). The detection of CC and SV attributes based uniquely on clue words found in the training set also underperformed wrt. other participants, both ranking last with differences of respectively -0.12 and -0.08 wrt. the best system.

The choice we made to process the Temporal Expressions with the Heideltime tool and specifically designed rules allowed us to achieve an accuracy of 0.839,

with a small δ of 0.025 wrt. the first system. The addition of the less represented *time* and *duration* classes was detrimental to this module.

The MaxEnt model we designed for negation identification performed well, achieving a 0.902 accuracy with a small δ of 0.067 wrt to the first system on this attribute (the maximum amplitude of accuracy on this attribute is of 0.207 between the first and the last system). However, the MaxEnt model we created for uncertainty identification obtained quite low results with an accuracy of 0.801, our system ranking last on this attribute. Given its similarity of design to the negation identification module and the very low precision and recall scores it obtained, we suspect this might be due to a bug in this module.

The detection of the DT attribute (temporal relation to document time) with an emphasis on the position of the disorder in the document structure (section type and relative position in document) performed on par with the best system. Its use of the document type as one of the features may have helped it perform well on the test set, which only contained discharge summaries, in contrast to the training set which included four types of documents. We have seen in further experiments on the training set that the use of similarity-based classifiers (kNN or SVM) instead of the decision tree might improve its results. Besides, it currently does not take into account verb tense, which can be expected to be an important clue for this attribute.

Finally, let us note that the accuracy scores obtained by the participants on the test corpus of this temporal relation task are the lowest among all attributes. They are much lower than those obtained on the training set (0.81 for our classifier in 10-fold cross-validation). They are also much lower that those obtained in the i2b2 2012 challenge on temporal relation detection (F-measures of the ten best systems in the 0.43–0.69 range) [18]. Our own work in the i2b2 2012 challenge [6] studied the relative recall of our classifiers. The temporal relations of i2b2 2012 that were closest to the DT attribute of the present task were those between an event and the admission (AD) or discharge (DD) date. For these two relations, we obtained F-measures, recalls and precisions of respectively (0.86, 0.80, 0.94) and (0.63, 0.51, 0.83) (see Figure 4 in [6], relations TIMEX3 EVENT DD HC and TIMEX3 EVENT AD HPI), also much higher than the scores for the DT attribute. However, the events in i2b2 2012 included more event types than only disorders, which may change the difficulty of the task.

5 Conclusion and Perspectives

We designed several systems to address the disease and disorder template filling task of ShARe/CLEF eHealth 2014. We chose the method to use (either rule-based or supervised approach) depending on the characteristics of each attribute: resource-based for attributes (e.g., BL) where a dictionary was an important component, rule-based where patterns were important (TE), based on supervised machine learning where the determination of the attribute value was based on distributions of features and relied on a study of their context (e.g., NI and DT).

While we achieved a high accuracy by using default values in the case of very unbalanced attributes, we consider that this is not satisfactory. A better study of contexts occurring near disorders should allow us to highlight clues that could be used either to produce rules or to train statistical models (taking into account the specific distribution of values of these attributes). The resource-based methods that we used probably need to be complemented with additional features to take better account of their context of occurrence. The supervised methods obtained high accuracies on the NI and UI attributes. The accuracy on the DT attribute was low for all participants, pointing at it as the hardest of all attributes: our system performed on par with the top system on this attribute, and we discussed directions to improve it further.

Acknowledgments

We acknowledge the Shared Annotated Resources (ShARe) project funded by the United States National Institutes of Health with grant number R01GM090187. This work was partly funded through project Accordys[4] funded by ANR under grant number ANR-12-CORD-0007-03.

References

1. Daniel Albright, Arrick Lanfranchi, Anwen Fredriksen, William F. 4th Styler, Colin Warner, Jena D. Hwang, Jinho D. Choi, Dmitriy Dligach, Rodney D. Nielsen, James Martin, Wayne Ward, Martha Palmer, and Guergana K. Savova. Towards comprehensive syntactic and semantic annotations of the clinical narrative. *J Am Med Inform Assoc*, 20(5):922–930, Sep-Oct 2013.
2. Adam L. Berger, Stephen Della Pietra, and Vincent J. Della Pietra. A maximum entropy approach to natural language processing. *Computational Linguistics*, 22(1):39–71, 1996.
3. Olivier Bodenreider. The Unified Medical Language System (UMLS): integrating biomedical terminology. *Nucleic Acid Res*, 32:D267–D270, 2004.
4. Noémie Elhadad, Wendy W. Chapman, Tim O'Gorman, Martha Palmer, and Guergana K. Savova. The ShARe schema for the syntactic and semantic annotation of clinical texts. 2014. Under Review.
5. Carol Friedman, Philip O. Alderson, John H. M. Austin, James J. Cimino, and Stephen B. Johnson. A general natural-language text processor for clinical radiology. *J Am Med Inform Assoc*, 1(2):161–174, 1994.
6. Cyril Grouin, Natalia Grabar, Thierry Hamon, Sophie Rosset, Xavier Tannier, and Pierre Zweigenbaum. Eventual situations for timeline extraction from clinical reports. *J Am Med Inform Assoc*, 20(5):820–827, Sep-Oct 2013. 2013 Apr 9. [Epub ahead of print].
7. Silviu Guiasu and Abe Shenitzer. The principle of maximum entropy. *The Mathematical Intelligence*, 7(1), 1985.

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

8. Mark A. Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. The WEKA data mining software: An update. *SIGKDD Explor Newsl*, 11(1), 2009.
9. Thierry Hamon and Natalia Grabar. Tuning heideltime for identifying time expressions in clinical texts in english and french. In *Proc of International Workshop on Health Text Mining and Information Analysis (LOUHI2014)*, pages 101–5, Gothenburg, Sweden, April 2014.
10. Thierry Hamon, Adeline Nazarenko, Thierry Poibeau, Sophie Aubin, and Julien Derivière. A robust linguistic platform for efficient and domain specific web content analysis. In *Proceedings of RIAO 2007*, Pittsburgh, USA, 2007. 15 pages.
11. Liadh Kelly, Lorraine Goeuriot, Gondy Leroy, Hanna Suominen, Tobias Schreck, Danielle L. Mowery, Sumithra Velupillai, Wendy W. Chapman, Guido Zuccon, and Joao Palotti. Overview of the ShARe/CLEF eHealth evaluation lab 2014. In *Proceedings of the ShARe/CLEF eHealth Evaluation Lab*. Springer-Verlag, 2014.
12. Thomas Lavergne, Olivier Cappé, and François Yvon. Practical very large scale CRFs. In *Proc of ACL*, pages 504–13, Uppsala, Sweden, July 2010.
13. Aurélie Névéol, Cyril Grouin, Jérémy Leixa, Sophie Rosset, and Pierre Zweigenbaum. The Quaero French medical corpus: A ressource for medical entity recognition and normalization. In *Proc BioTextM*, Reykjavik, Iceland, 2014.
14. Kirk Roberts, Bryan Rink, Sanda M. Harabagiu, Richard H. Scheuermann, Seth Toomay, Travis Browning, Teresa Bosler, and Ronald Peshock. A machine learning approach for identifying anatomical locations of actionable findings in radiology reports. In *AMIA Annu Symp Proc*, volume 2012, pages 779–788, 2012.
15. Mohammed Saeed, Mauricio Villarroel, Andrew T. Reisner, Gari Clifford, Li-Wei Lehman, George B. Moody, Thomas Heldt, Tin H. Kyaw, Benjamin E. Moody, and Roger G. Mark. Multiparameter intelligent monitoring in intensive care II (MIMIC-II): A public-access ICU database. *Clin Care Med*, 39:952–960, 2011.
16. Naomi Sager, Carol Friedman, and Margaret S. Lyman, editors. *Medical Language Processing: Computer Management of Narrative Data*. Addison Wesley, Reading, MA, 1987.
17. Jannik Strötgen and Michael Gertz. Temporal tagging on different domains: Challenges, strategies, and gold standards. In *Proc of LREC*, pages 3746–3753, 2012.
18. Weiyi Sun, Anna Rumshisky, and Özlem Uzuner. Evaluating temporal relations in clinical text: 2012 i2b2 challenge overview. *J Am Med Inform Assoc*, 20(5):806–813, Sep-Oct 2013.
19. Hanna Suominen, Sanna Salanterä, Sumithra Velupillai, Wendy W. Chapman, Guergana K. Savova, Noémie Elhadad, Sameer Pradhan, Brett R. South, Danielle L. Mowery, Gareth J. F. Jones, Johannes Leveling, Liadh Kelly, Lorraine Goeuriot, David Martinez, and Guido Zuccon. Overview of the ShARe/CLEF eHealth evaluation lab 2013. In *Proceedings of CLEF 2013*, Lecture Notes in Computer Science, Berlin Heidelberg, 2013. Springer.
20. 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 *Proc of Advances in Informatics – 10th Panhellenic Conference on Informatics*, LNCS 3746, pages 382–92, 2005.
21. Özlem Uzuner, Brett R. South, Shuying 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, Sep-Oct 2011. Epub 2011 Jun 16.
22. Pierre Zweigenbaum. MENELAS: an access system for medical records using natural language. *Computer Methods and Programs in Biomedicine*, 45:117–120, 1994.