Journal of Biomedical Informatics 46 (2013) 425–435

Contents lists available at SciVerse ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

An enhanced CRFs-based system for information extraction from radiology reports [☆]

Andrea Esuli, Diego Marcheggiani, Fabrizio Sebastiani [*]

Istituto di Scienza e Tecnologie dell'Informazione, Consiglio Nazionale delle Ricerche, 56124 Pisa, Italy

ARTICLE INFO

Article history:
Received 21 July 2012
Accepted 28 January 2013
Available online 11 February 2013

Keywords:
Information extraction
Concept extraction
Radiology reports
Medical reports
Clinical narratives
Machine learning
Conditional random fields

ABSTRACT

We discuss the problem of performing information extraction from free-text radiology reports via supervised learning. In this task, segments of text (not necessarily coinciding with entire sentences, and possibly crossing sentence boundaries) need to be annotated with tags representing concepts of interest in the radiological domain. In this paper we present two novel approaches to IE for radiology reports: (i) a cascaded, two-stage method based on pipelining two taggers generated via the well known linear-chain conditional random fields (LC-CRFs) learner and (ii) a confidence-weighted ensemble method that combines standard LC-CRFs and the proposed two-stage method. We also report on the use of "positional features", a novel type of feature intended to aid in the automatic annotation of texts in which the instances of a given concept may be hypothesized to systematically occur in specific areas of the text. We present experiments on a dataset of mammography reports in which the proposed ensemble is shown to outperform a traditional, single-stage CRFs system in two different, applicatively interesting scenarios.

1. Introduction

Information Extraction (IE – sometimes also referred to as *concept extraction*) is the discipline concerned with the extraction of natural language expressions from free text, where these expressions instantiate concepts of interest in a given domain (see e.g., [1,2]). An application of great interest is extracting information from free-text clinical narratives, such as medical reports, admission/ discharge summaries, and clinical notes in general. These narratives are unstructured in nature, since they are written in free text by medical personnel, and because of their informal nature are particularly difficult to handle automatically. Medical reports and clinical narratives are characterized by informal language, and are usually fraught with abbreviations (sometimes idiosyncratic to the specific hospital or department where they originated), ungrammatical language, acronyms (even non-standard ones), and typos; this is due to the fact that clinical narratives are often the result of hasty compilation, or dictation, or the application of speech recognition technology. As a result, the correct identification of expressions (e.g., named entities, phrases, sentences, or other) that instantiate concepts of interest (such as drugs, drug dosages, pathologies, treatments, and the like) is more difficult than, say, in the domain of medical literature (e.g., books, abstracts, scientific articles), where language is usually tighter.

Nonetheless, performing information extraction on these narratives would be extremely beneficial, since extracting key data and converting them into a structured (e.g., tabular) format would greatly contribute towards endowing patients with truly actionable electronic medical records. These records, aside from improving interoperability among different medical information systems, could be used in a variety of applications, from patient care, to decision support, to epidemiology and clinical studies. Of particular interest is the fact that automatically spotting personal information in narratives may be used for their automatic de-identification (i.e., anonymization) [3], an important task given the confidential nature of clinical narratives.

The present paper describes our efforts towards automatically extracting information from radiology reports. There are two main approaches to designing such a system. One is the rule-based approach, which consists in manually writing a set of pattern-matching rules which relate natural language patterns with the concepts to be extracted from the text. This approach, while potentially effective, is very costly, since it requires a lot of human effort for writing the rules, which need to be collaboratively written by a domain expert – say, an expert radiologist – and a natural language engineer. We have followed the alternative approach, which is based on supervised machine learning. According to this approach, a general–purpose learning software learns from examples to relate natural language patterns with the concepts of interest, using as examples a set of manually annotated free texts, i.e., texts in which the instances of the concepts of

[☆] The order in which the authors are listed is purely alphabetical; each author has given an equally important contribution to this work.

[*] Corresponding author. Fax: +39 050 3153464.
E-mail address: fabrizio.sebastiani@isti.cnr.it (F. Sebastiani).

1532-0464/$ - see front matter © 2013 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.jbi.2013.01.006

interest have been marked as such by a domain expert. The advantage of this approach is that the human effort required for annotating the texts needed for training the system is much smaller, and requires less expertise, than the one needed for manually writing the extraction rules.

Nowadays, the dominant machine learning method for IE from clinical narratives is *conditional random fields* [4–6], a method explicitly devised for *sequence learning*, i.e., for learning to classify items that naturally occur in sequences (the words that make up the clinical narratives obviously have this property). Most authors that have participated in the recent i2b2 challenges devoted to IE from clinical narratives (see [7,8]) have indeed used CRFs [9–12], often in their simple "linear chain" form (LC-CRFs). In this paper we present two novel approaches to using LC-CRFs for clinical text IE: (i) a cascaded, two-stage method based on pipelining two LC-CRFs systems, one that analyses text at the clause level and a second one that analyses it at the token (word) level and (ii) a confidence-weighted ensemble method that combines a standard, token-level LC-CRFs and the proposed cascaded method.

When input to a CRFs learner, each token (i.e., word instance) occurring in the narrative must be represented as a vector of features. Typical features that are used in CRFs for clinical IE are the word the token is an instance of, its morphological root, its part of speech, its prefixes and suffixes, and other information that can directly be extracted from the text; the features of the surrounding tokens are also usually added to the vector representing a token. If specialized lexical resources are available (e.g., the UMLS metathesaurus), entries from these resources can also be usefully added to the representation of a token whenever the token, or a larger sequence of tokens that includes it, is recognized as standing in certain relationships with those entries. In this paper we report on the additional use of "positional features", a novel type of features intended to aid in the automatic annotation of texts in which the instances of a given concept tend to systematically occur in specific areas of the text (e.g., towards the beginning of the text, towards the middle of the text, etc.).

We experimentally validate our newly introduced methods on a dataset of 500 mammography reports written in Italian. Two applicatively interesting scenarios are tested. The first scenario simulates the existence of a single human annotator who provides the training data and whose classification decisions are to be taken as the gold standard. The second scenario simulates the existence of multiple human annotators; a mix of training examples from each annotator are thus used, and a mix of the classification decisions of each of them is used as the gold standard. This also represents a novelty in the literature on clinical IE, since this literature has never (to the best of our knowledge) taken up the issue of distinguishing, and separately addressing, these two applicative scenarios.

The results of our experiments show that our two proposed LC-CRFs systems outperform, in both scenarios, a baseline system consisting of linear-chain CRFs in the pure form, and show that in most cases the performance of our systems even exceeds human performance, as measured by inter-coder agreement. The results obtained from the use of positional features are instead less conclusive.

The rest of the paper is organized as follows. In Section 2 we describe the LC-CRFs system and the set of features that constitute our baseline. In Section 3 we turn to describing our proposed novel methods: the cascaded LC-CRFs system (Section 3.1), the ensemble of LC-CRFs learners (Section 3.2), and the positional features (Section 3.3). Section 4 is devoted to describing our experiments and commenting on their results. Section 5 discusses related work, while Section 6 concludes, pointing at avenues for future research.

2. A baseline IE system for radiology reports

2.1. Preliminaries and notation

Let $\mathbf{X}$ be a set of texts. Let a text $\mathbf{x} \in \mathbf{X}$ consist of a vector $\mathbf{x} = \langle x_1, \ldots, x_{|\mathbf{x}|} \rangle$, where each x_t is a *token* (i.e., a word occurrence), $|\mathbf{x}|$ denotes the dimensionality of vector $\mathbf{x}$ (in this case: the length of the text), and x_{t_1} occurs before x_{t_2} in the text (noted $x_{t_1} \preceq x_{t_2}$) if and only if $t_1 \leqslant t_2$. Let $C = \{c_1, \ldots, c_m\}$ be a predefined set of *tags* (a.k.a. *concepts*, or *markables*), or *tagset*. We will take *information extraction* (IE) to be the task of determining, for each $\mathbf{x} \in \mathbf{X}$ and for each $c_r \in C$, a vector $\mathbf{y}_r = \langle y_{r1}, \ldots, y_{r|\mathbf{x}|} \rangle$ of labels $y_{rt} \in \{c_r, \bar{c}_r\}$, which indicates which tokens in the text are labelled with tag c_r. Since each $c_r \in C$ is dealt with independently of the other tags in C, we will hereafter drop the r subscript and treat IE as the *binary* task of determining, given text $\mathbf{x}$ and given tag c, a vector $\mathbf{y} = \langle y_1, \ldots, y_{|\mathbf{x}|} \rangle$ of labels $y_t \in \{c, \bar{c}\}$. We assume each token x_t to be itself represented by a vector $\mathbf{x}_t$ of $\Omega = |\mathbf{x}_t|$ features.

Tokens labelled with a tag c usually come in coherent sequences, or "segments". Hereafter, a *segment* σ of text $\mathbf{x}$ for tag c will be a pair (x_{t_1}, x_{t_2}) consisting of a start token x_{t_1} and an end token x_{t_2} such that (i) $x_{t_1} \preceq x_{t_2}$, (ii) all tokens $x_{t_1} \preceq x_t \preceq x_{t_2}$ are labelled with tag c, and (iii) the token that immediately precedes x_{t_1} and the one that immediately follows x_{t_2} are *not* tagged with tag c. For example, in the excerpt

> `Da ambo i lati si apprezzano` [PAE] `formazioni nodulari solide` [/PAE], `a margini sostanzialmente regolari.`

the markers [PAE] and [/PAE] are meant to indicate that all the tokens between them are to be understood as tagged with the `PresenzaAssenzaEnhancement` (PAE) tag. In this case, (`formazioni`, `solide`) is a segment for the PAE tag. In general, a text $\mathbf{x}$ may contain zero, one, or several segments for tag c.

2.2. A baseline set of features

As a baseline learning algorithm we have used *linear-chain conditional random fields* (LC-CRFs – [4–6]), in Charles Sutton's GRMM implementation.[1] LC-CRFs is a class of supervised learning algorithms explicitly devised for *sequence labelling*, i.e., for learning to label items that naturally occur in sequences and such that the label of an item may depend on the features and/or on the labels of other items that precede or follow it in the sequence (which is indeed the case for the tokens in a text). See Appendix A for a mathematical explanation of LC-CRFs.

As mentioned in Appendix A, a CRFs-based learner needs each token x_t to be represented by a vector $\mathbf{x}_t$ of features. It is thus a key design decision to select which linguistic characteristics of x_t, or of the tokens that immediately precede or follow it, should be represented in the feature vector $\mathbf{x}_t$.

In this work we have used a base set of features which includes:

1. One feature representing the word of which the token is an instance.
2. One feature representing its stem.
3. One feature representing its part of speech.
4. Eight features representing its prefixes and suffixes (the first and the last n characters of the token, with $n = 1, 2, 3, 4$).
5. One feature representing information on token capitalization, i.e., whether the token is all uppercase, all lowercase, first letter uppercase, or mixed case.

Vector $\mathbf{x}_t$ includes the above information for each of x_{t-1}, x_t and x_{t+1}, for a total of 36 features.

[1] `http://mallet.cs.umass.edu/grmm/`.

All of these features are fairly standard in several instances of the information extraction task, including e.g., named entity recognition. The IE system we will use as a baseline in the experiments of Section 4 will thus consists of the LC-CRFs learning system discussed in Appendix A using as input the text represented via the features discussed in the present section.

3. An enhanced IE system for radiology reports

3.1. A two-stage CRFs-based IE system

In medical reports it is often the case that the segments of interest are not random chunks of text, but span several clauses (or even several sentences). A given tag may be expected to be instantiated by an expression that spans several clauses, maybe cutting across sentence borders. This suggests the adoption of a cascaded, two-stage tagging scheme, in which

1. The first stage consists of tagging entire clauses; this acts as a coarse-grained filter, with the goal of removing from consideration clauses that are evidently irrelevant to the tag of interest.
2. The second stage consists of tagging individual tokens belonging to the clauses that have been attributed the tag in the first stage; this acts as a fine-grained tagger, with the goal of examining in detail the clauses that the previous phase has deemed of potential interest.

For the purpose of this paper we heuristically define a *clause* as a set of tokens delimited both to the right and to the left by a punctuation symbol in the set {comma, period, colon, semicolon, question mark, exclamation mark}, and such that no punctuation symbol from this set appears inside it. For example, in the excerpt

> `Da ambo i lati si apprezzano` [PAE] `formazioni nodulari solide` [/PAE], `a margini sostanzialmente regolari.`

already quoted in Section 2.1, the clauses are `Da ambo i lati si apprezzano formazioni nodulari solide` and `a margini sostanzialmente regolari`.

As for the first stage, we implement it via a LC-CRFs tagger that, unlike the one described in Appendix A, has the formulation

$$p(\mathbf{y}|\mathbf{s}) = \frac{1}{Z(\mathbf{s})}\prod_{d=1}^{D}\Psi_u(y_d, \mathbf{s}_d)\prod_{d=1}^{D-1}\Psi_b(y_d, y_{d+1}, \mathbf{s}_d) \qquad (1)$$

Here, the sequence of tokens $\mathbf{x} = \langle x_1, \ldots, x_{|\mathbf{x}|}\rangle$ has been replaced with a sequence of clauses $\mathbf{s} = \langle s_1, \ldots, s_{|\mathbf{s}|}\rangle$, represented by feature vectors $\mathbf{s}_1, \ldots, \mathbf{s}_{|\mathbf{s}|}$, with all the equations of Appendix A rewritten accordingly.

We see the first stage as a sort of *text classification* phase, where entire texts (in this case: clauses) need to be classified, using the tags in our tagset as the classes. Accordingly, we opt for a "bag of words" (and word bigrams) representation, thus using as only features the words of which the tokens in the clause are instances, and their bigrams. As a consequence, the $\mathbf{s}_t$ feature vector is the traditional sparse vector of text classification, whose dimensionality is the number of unique words and bigrams that occur at least once in the training set, and where 0 and 1 represent absence and presence of the word or bigram in the clause. However, note that we are not classifying a clause in isolation of the clauses that precede and follow it. The very fact that we are using a LC-CRFs approach, and that this involves the feature functions of Eq. (A.5), means that the tag that will be attributed to a given clause will also depend on the tag that is probabilistically attributed to the clause the follows it.

Quite obviously, we also implement the second stage via the standard LC-CRFs system described in Section 2, with the difference that clauses, instead of entire texts, are the object of tagging. Only the clauses that have passed through the filter of the first-stage system are tagged (at the token level). The evaluation of the cascaded system is performed by considering the token-level representation of the clauses discarded in the first stage as labelled with tag $\bar{c}$.

3.2. A confidence-weighted ensemble of CRFs-based IE systems

The two-stage CRFs described in Section 3.1 trades precision for recall. In fact, it can be expected to have better precision than the single-stage system, since all the clauses that are ruled out from consideration in the first stage may contain tokens that are erroneously attributed the tag by the single-stage system; by removing these clauses from consideration, a number of potential false positives become true negatives, thereby improving precision. However, by the same argument, the cascaded, two-stage system may be expected to have worse recall then the single-stage system, since the ruled-out clauses may contain tokens to which the single-stage system would have *correctly* attributed the tag; this generates false negatives that would otherwise have been true positives.

In sum, better precision but worse recall may be expected from going two-stage. Whether the two-stage system is better than the single-stage one thus comes down to understanding whether, by which amount, and for which tags, the improvement in precision outweigh the deterioration in recall.

One may want to try to obtain the best of both worlds by having a committee (or ensemble) of two taggers (a single-stage one and a two-stage one), where the final decision is also based on how confident each of the two taggers is in tagging a given token.

More specifically, let us note that the Viterbi algorithm used in both the single-stage and the cascaded, two-stage systems for maximizing $p(\mathbf{y}|\mathbf{x})$, also maximizes, as an intermediate result, the conditional probabilities $p(y_t|\mathbf{x})$ of the individual tokens x_t. Let p_{ss}(-$y_t|\mathbf{x}$) and $p_{ts}(y_t|\mathbf{x})$ denote these probabilities as maximized by the single-stage and the two-stage methods, respectively. We define the conditional probability $p_{en}(y_t|\mathbf{x})$ of the individual token x_t as maximized by our ensemble, as the average of the conditional probabilities computed by the single-stage and the two-stage methods, i.e.,

$$p_{en}(y_t|\mathbf{x}) = \frac{1}{2}(p_{ss}(y_t|\mathbf{x}) + p_{ts}(y_t|\mathbf{x})) \qquad (2)$$

Since the conditional probability $p(y_t|\mathbf{x})$ may obviously be interpreted as the system's confidence in the fact that the label of $\mathbf{x}_t$ is y_t, Eq. (2) formalizes a *confidence-weighted ensemble* [13]. Note that $p_{ts}(y_t|\mathbf{x})$ is obtained by the cascaded, two-stage tagger in the second stage whenever the clause that contains x_t has not been ruled out in the first stage. Otherwise, $p_{ts}(y_t|\mathbf{x})$ is in fact the probability that the first-stage tagger has attributed y_t to the entire clause which contains x_t, since all of the tokens in that clause obtain the same probability.

Therefore, our ensemble-based method consists of computing

$$\mathbf{y}_t^* = \arg\max_{y_t} p_{en}(y_t|\mathbf{x}) \qquad (3)$$

for each token x_t.

3.3. Positional features

As for representing tokens x_t via feature vectors $\mathbf{x}_t$, we think that there are margins of improvement over the choices discussed in Section 2.2. In particular, we observe that medical reports are often written according to a fairly standard pattern, according to which some tags are instantiated earlier on in the report, while some others are instantiated later. For instance, it is reasonable to suppose that clinical observations are presented at the beginning of the report, while a diagnosis and a prognosis are discussed later on.

In order to capture these recurring patterns, we introduce *positional features* that model the position of a token in the text. For instance, we might want to introduce a 4-ary feature that indicates whether token x_t occurs in the 1st quarter of the text, or in the 2nd, or in the 3rd, or in the 4th. In this way, if a given tag tends to be instantiated, say, in the 1st quarter of the report, this tendency will be detected during training, and will be brought to bear in the test phase, e.g., by using the fact that test token x_t occurs in the 1st quarter of the text as contributing evidence that x_t should be assigned the tag. In general, our positional features are k-ary features which, assuming that the text is divided into k consecutive, equal-sized parts, indicate in which of the k parts the token occurs.

In our experiments we use all positional features for $k = 2,3,4,5$.

4. Experiments

4.1. The evaluation measure

As the evaluation measure we use the token-level variant (proposed in [14]) of the well-known F_1 measure, according to which a tagger is evaluated on an event space consisting of all tokens in the text. In other words, each token x_t (rather than each segment, as in the traditional "segmentation F-score" model [15]) counts as a true positive, true negative, false positive, or false negative for a given tag c_r, depending on whether x_t belongs to c_r or not in the predicted annotation and in the true annotation. As argued in [14], this model has the advantage that it credits a system for partial success (i.e., non-null overlap between a predicted segment and a true segment for the same tag), and that it penalizes both overtagging and undertagging.

As is well-known, F_1 combines the contributions of *precision* (π) and *recall* (ρ), and is defined as

$$F_1 = \frac{2\pi\rho}{\pi + \rho} = \frac{2TP}{2TP + FP + FN}$$

where TP, FP, and FN stand for the numbers of true positives, false positives, and false negatives, respectively. Note that F_1 is undefined when $TP = FP = FN = 0$; in this case we take F_1 to equal 1, since the tagger has correctly tagged all tokens as negative.

We compute F_1 across the entire test set, i.e., we generate a single contingency table by putting together all tokens in the test set, irrespectively of the text they belong to. We then compute both *microaveraged* F_1 (denoted by F_1^{μ}) and *macroaveraged* F_1 $\left(F_1^M\right)$. F_1^{μ} is obtained by (i) computing the tag-specific values TP_r, FP_r and FN_r, (ii) obtaining TP as the sum of the TP_r's (same for FP and FN), and then (iii) applying the $F_1 = \frac{2TP}{2TP+FP+FN}$ formula. F_1^M is obtained by first computing the tag-specific F_1 values and then averaging them across the c_r's.

An advantage of using F_1 as the evaluation measure is that it is symmetric, i.e., its values do not change if one switches the roles of the human annotator (i.e., the gold standard) and the automatic annotator (i.e., the system). This means that F_1 can also be used as a measure of agreement between any two annotators, regardless of whether they are human or machine, since it does not require one to specify who among the two is the gold standard against which the other needs to be checked. For this reason, in the following section we will use F_1 both (a) to measure the agreement between our system and a human annotator *and* (b) to measure the agreement between two human annotators.

Table 1
Per-tag statistics for the Umbertol(RadRep) dataset. Columns 2 and 3 indicate the number $NT(i)$ of tokens and the number $NS(i)$ of segments contained in the dataset for the given tag, Column 4 the number $ND(i)$ of documents with at least one segment for the given tag, Column 5 the average number $ANS(i)$ of segments per document for the given tag, Column 6 the average segment length $ASL(i)$ for the given tag (segment length is the number of tokens contained in it).

	$NT(i)$	$NS(i)$	$ND(i)$	$ANS(i)$	$ASL(i)$
BI-RADS (BIR)	1544	400	293	0.65	3.86
InformazioniTecniche (ITE)	24,301	615	614	0.99	39.51
IndicazioniEsame (IES)	5159	475	469	0.77	10.86
TerapieFollowup (TFU)	7137	523	458	0.84	13.65
DescrizioneEnhancement (DEE)	19,795	803	440	1.30	24.65
PresenzaAssenzaEnhancement (PAE)	8461	588	439	1.94	7.04
EsitiChirurgici (ECH)	2068	230	203	0.37	8.99
DescrizioneProtesi (DEP)	2532	72	61	0.12	35.17
LinfonodiLocoregionali (LLO)	6602	537	509	0.87	12.29
Average	8622	471	387	0.87	17.33

4.2. The dataset

The dataset we have used to test the ideas presented in Section 3 (hereafter called the Umbertol(RadRep) dataset) consists of a set of 500 free-text mammography reports written (in Italian) by medical personnel of the Istituto di Radiologia of Policlinico Umberto I, Roma, Italy. The number of different radiologists who authored the reports is not known. The length of the reports ranges between 67 and 537 words; the mean and the variance of such length is 199 and 8786, respectively.

The reports have been subsequently annotated by two equally expert radiologists from the same institute; 191 reports have been annotated by annotator 1 (A1) only, 190 reports have been annotated by annotator 2 (A2) only, and 119 reports have been annotated independently by A1 and A2. From now on we will call these sets `A1-only`, `A2-only` and `Both`, respectively; `Both(1)` will identify the `Both` set as annotated by A1, and `Both(2)` will identify the `Both` set as annotated by A2. The annotation activity was preceded by an alignment phase, in which A1 and A2 jointly annotated 20 reports (not included in this dataset) in order to align their understanding of the meaning of the tags.

The tagset is formed by 9 tags: BI-RADS (hereafter shortened as BIR),[2] InformazioniTecniche (ITE – "Technical Info"), IndicazioniEsame (IES – "Indications obtained from the Exam"), TerapieFollowup (TFU – "Followup Therapies"), DescrizioneEnhancement (DEE – "Description of Enhancement"), PresenzaAssenzaEnhancement (PAE – "Presence/Absence of Enhancements"), EsitiChirurgici (ECH – "Outcomes of Surgery"), DescrizioneProtesi (DEP – "Prosthesis Description"), and LinfonodiLocoregionali (LLO – "Locoregional Lymph Nodes"). On average, there are 0.87 segments for each tag in a given report, and the average segment length is 17.33 words. Table 1 reports additional statistics on the dataset; Fig. 1 illustrates, in histogram form, the variability in report length across the dataset; Fig. 2 illustrates a sample, manually annotated report from the Umbertol(RadRep) dataset.

A closer inspection of this dataset reveals that

- It is not the case that that for each tag $c_r \in C$ there is at least one segment σ_{rj} in each report. For example, the only tags that are instantiated in Report 114 are BIR, PAE, ITE, and TFU.
- Segments for different tags may overlap, i.e., the same token x_t may belong to two or more segments each pertaining to a different tag. For example, in Report 452 all the tokens in the

[2] BI-RADS indicates the mammography assessment categories from the Breast Imaging-Reporting and Data System published by the American College of Radiology (ACR). These are standardized numerical codes typically assigned by a radiologist after interpreting a mammogram, which allow for concise and unambiguous understanding of patient records among doctors.

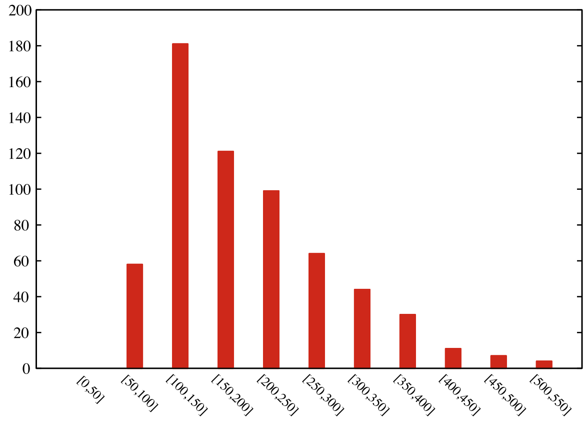


Fig. 1. Distribution of report lengths: the X axis represents ranges of lengths, while the Y axis represents the number of reports whose length falls in the given range.

sequence `in paziente con pregressa QUART sinistra mastoplastica additiva bilaterale` are tagged with both tags ECH and IES.

- Two or more segments for the same tag c_r may be present in the same report. For instance, in Report 180 the two non-overlapping segments `un’area di potenziamento dopo contrasto` and `alcuni millimetrici foci di potenziamento` are both tagged with tag PAE.
- The “order of appearance” of the tags in a report is not fixed, i.e., given tags c' and c'' it is *not* the case that segments for tag c' always precede the segments for tag c''. For example, in Report 452 tag PAE appears after tag DEE, while in Report 180 it appears before it.

4.3. The experimental setting

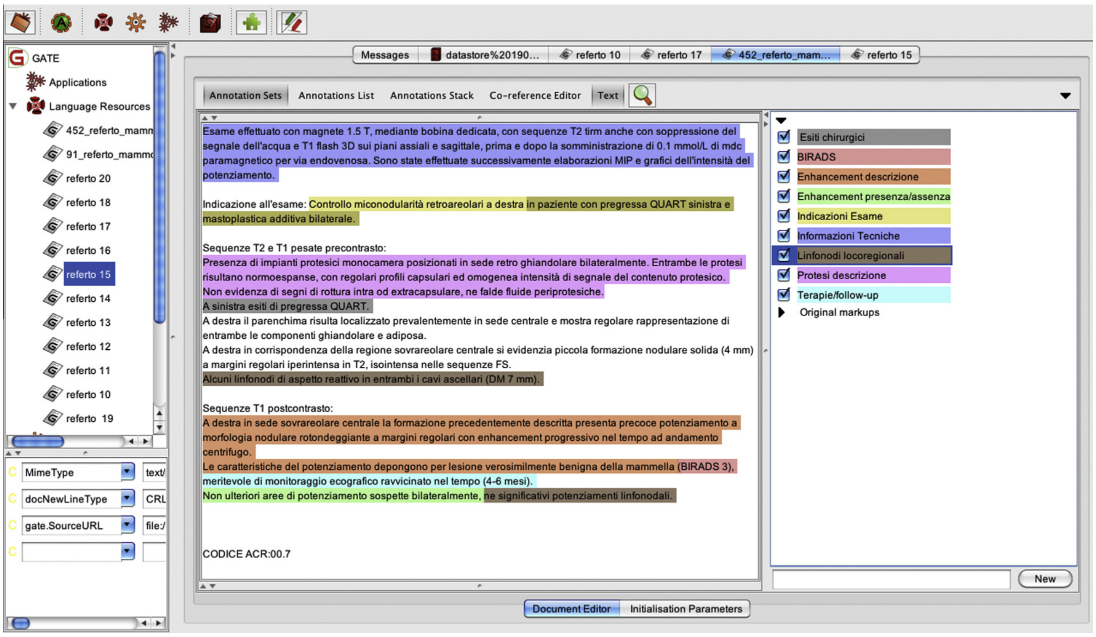
We have tested three different learning systems. The first is the LC-CRFs system described in Appendix A, which we use as the baseline, while the second and the third are our cascaded, two-stage LC-CRFs system and our ensemble of taggers described in Sections 3.1 and 3.2, respectively. In combination with each of the three learning systems we have also tested two different feature representations, one consisting of the baseline features described in Section 2.2, and one also including the positional features discussed in Section 3.3. Concerning these latter, note that in the first stage of the cascaded system (when used by itself and when used in the ensemble system) we have used positional features *for the clauses*, defined as the positional feature of their middle tokens.

We have optimized the regularization parameter δ of Eq. (A.8) individually for both the baseline system of Appendix A and the two-stage system of Section 3.1, and individually for each experiment reported in this paper. We have always carried out this optimization on a held-out validation set consisting of a randomly selected 10% of the original training set; after parameter optimization, the tagger has then been retrained on the entire training set. As a POS tagger we have used the one contained in the TextPro system [17].

We have run two main sets of experiments, each simulating a different operational scenario.

The first such set of experiments simulates a *single-annotator scenario*, i.e., an operational situation in which the organization has a single person in charge of annotating the reports. Simulating this scenario thus means using training and test documents annotated by the same person, since it is this person who would annotate the training data to be fed to a learning algorithm, and it is this person who would judge the accuracy of the automatically tagged data. We have thus run (i) an experiment in which we train the system on `A1-only` and test it on `Both(1)`, and (ii) an experiment in which we train the system on `A2-only` and test it on `Both(2)`. Since there is not much value in presenting the results of these two experiments separately, in Table 2 we report accuracy figures obtained by averaging the results of the two experiments (or, equivalently, by merging, for each tag in the tagset, the two contingency tables obtained in the two experiments into a single contingency table and evaluating the results).

The second set of experiments simulates instead a *multiple-annotator scenario*, i.e., one in which there are several, equally



Esame effettuato con magnete 1.5 T, mediante bobina dedicata, con sequenze T2 tirm anche con soppressione del segnale dell'acqua e T1 flash 3D sui piani assiali e sagittale, prima e dopo la somministrazione di 0.1 mmol/L di mdc paramagnetico per via endovenosa. Sono state effettuate successivamente elaborazioni MIP e grafici dell'intensità del potenziamento.

Indicazione all'esame: Controllo miconodularità retroareolari a destra in paziente con pregressa QUART sinistra e mastoplastica additiva bilaterale.

Sequenze T2 e T1 pesate precontrasto:
Presenza di impianti protesici monocamera posizionati in sede retro ghiandolare bilateralmente. Entrambe le protesi risultano normoespanse, con regolari profili capsulari ed omogenea intensità di segnale del contenuto protesico. Non evidenza di segni di rottura intra od extracapsulare, ne falde fluide periprotesiche.
A sinistra esiti di pregressa QUART.
A destra il parenchima risulta localizzato prevalentemente in sede centrale e mostra regolare rappresentazione di entrambe le componenti ghiandolare e adiposa.
A destra in corrispondenza della regione sovrareolare centrale si evidenzia piccola formazione nodulare solida (4 mm) a margini regolari iperintensa in T2, isointensa nelle sequenze FS.
Alcuni linfonodi di aspetto reattivo in entrambi i cavi ascellari (DM 7 mm).

Sequenze T1 postcontrasto:
A destra in sede sovrareolare centrale la formazione precedentemente descritta presenta precoce potenziamento a morfologia nodulare rotondeggiante a margini regolari con enhancement progressivo nel tempo ad andamento centrifugo.
Le caratteristiche del potenziamento depongono per lesione verosimilmente benigna della mammella (BIRADS 3), meritevole di monitoraggio ecografico ravvicinato nel tempo (4-6 mesi).
Non ulteriori aree di potenziamento sospette bilateralmente, ne significativi potenziamenti linfonodali.

CODICE ACR:00.7

Fig. 2. A screenshot displaying a mammographic report automatically annotated according to the nine concepts of interest. The screenshot depicts the interface of the GATE system [16], which the two human annotators have used for manually annotating the reports.

Table 2
Results of the single-annotator experiments. Results are averages across two experiments, (i) train the system on A1-only and test it on Both(1), and (ii) train the system on A2-only and test it on Both(2). The first row reports the inter-annotator agreement values for A1 and A2 as measured on Both.

	BIR	ITE	IES	TFU	DEE	PAE	ECH	DEP	LLO	F_1^μ	F_1^M
A1 vs. A2	0.492	0.993	0.778	0.815	0.660	0.694	0.574	0.912	0.884	0.812	0.756
Baseline	0.583	0.987	0.772	0.872	0.745	0.717	0.667	0.812	0.848	0.846	0.778
Two-Stage	0.582	0.993	0.792	0.845	0.746	0.694	0.666	0.772	0.852	0.873	0.771
Ensemble	0.601	0.991	0.814	0.873	0.772	0.730	0.679	0.823	0.853	0.858	0.793
Baseline + PFs	0.537	0.987	0.810	0.872	0.742	0.717	0.614	0.803	0.872	0.848	0.773
Two-Stage + PFs	0.480	0.991	0.812	0.825	0.736	0.691	0.664	0.890	0.853	0.870	0.771
Ensemble + PFs	0.551	0.991	0.821	0.865	0.775	0.719	0.673	0.881	0.861	0.859	0.793

trustworthy persons in the organization who are in charge of annotation. Simulating the latter scenario thus means using a mix of training documents from different annotators (since it is not a single person who would annotate the training data to be fed to the learning algorithm) *and* a mix of test documents from different annotators (since it is not according to the judgment of a single person that the accuracy of the automatically tagged data should be measured). We have thus run an experiment in which we train the system on the union of A1-only and A2-only, and test it on the union of Both(1) and Both(2). Note that the union of Both(1) and Both(2) contains "contradictory information", since it contains two copies of each report in Both, and these two copies may contain annotations that contradict each other. This is very much in keeping with the fact that there are multiple, equally trustworthy annotators in the scenario we are simulating. The only practical consequence is that, in the presence of even the slight disagreement between Both(1) and Both(2), a value of $F_1 = 1$ is unattainable even in theory. However, this is not problematic for our experiments, and the value of inter-coder agreement of A1 and A2 as measured on Both via F_1 may be assumed as the "reasonable" accuracy upper bound that an automatic system cannot be expected to exceed. Table 3 reports the results obtained in this experiment.

Note that the two experiments are different in terms of both (i) training data *consistency* (since in the single-annotator experiments the annotations are all by the same person, hence are likely more consistent than in the multiple-annotator experiments), and (ii) training data *quantity* (since in the multiple-annotator experiment there are twice as many training data than in the single-annotator ones). This means that a comparison between the accuracy values obtained by a given system in these two sets of experiments will allow us to determine whether, for training data, consistency is a more important parameter than sheer quantity.

Note also that in both experiments the test documents are those in Both, i.e., those that have been annotated independently by both A1 and A2. In each such experiment we are thus in a position to compare the results of the experiments with the value of inter-annotator agreement obtained by A1 and A2 on the very same data. We are thus able to directly *compare human accuracy with machine accuracy* in a meaningful way, since, as argued in Section 4.1, F_1 can be used to measure the agreement between any two annotators, be them human or machine. More precisely,

- In the single-annotator experiment in which we train the system on A1-only and test it on Both(1), comparing the results of the experiment with the value of inter-coder agreement means comparing, using Both(1) as the common testbed (i.e., using A1 as the judge), an automatic system trained on A1-only against human annotator A2 (analogously for the experiment in which we train on A2-only and test on Both(2)).
- In the multiple-annotator experiment in which we train the system on the union of A1-only and A2-only, and test it on the union of Both(1) and Both(2), comparing the results of the experiment with the value of inter-coder agreement means comparing, using the union of Both(1) and Both(2) as the common testbed (i.e., using a mix of A1 and A2 as the judges), an automatic system trained on the union of A1-only and A2-only against a mix of A1 and A2.

We have also tested the statistical significance of the improvements obtained by our proposed methods with respect to the baseline. Specifically, we have compared the F_1 values as measured on each report, for each tag and for each pair of methods we compare. Given that we can exactly pair the F_1 value for a given method to the corresponding F_1 value for the other method, we have used a one-tailed paired *t*-test; we have used the one-tailed version of the test because we want to determine the significance of the improvement observed for the novel method with respect to the baseline method.

4.4. Results

4.4.1. Single-annotator results

The first observation we can make from Table 2 is that both the F_1^μ and F_1^M values obtained by all the tested machine-learned systems are clearly higher than the corresponding inter-coder agreement values. This indicates that, in our single-annotator scenario experiments, all our systems have proven to be even more accurate than humans. The difference between human performance and system performance is actually quite varied across the nine tags. In general, it seems that the system has a more uniform performance across the tags than the humans, who instead perform very well on some tags (e.g., ITE and DEP) and very bad on others (e.g., BIR and ECH).[3]

The results of the systems that make use of positional features (indicated as "+ PFs" in Table 2) are not statistically significantly different from the analogous systems that make no use of such features. A more fine-grained analysis does not seem to reveal any discernible pattern concerning which tags benefit from positional features and which do not. In fact, it might seem plausible to hypothesize that the tags whose instances tend to be more heavily concentrated in a specific quantile of the text (such as, e.g., tag ITE, which is heavily concentrated in the 1st and 2nd quintiles – see Fig. 3)) are the ones who benefit most from positional features. This hypothesis is not confirmed by the results. For instance, for the above-mentioned tag ITE, $F_1 = .987$ for both versions with and without positional features of the baseline system. Two tags such as IES and ECH which, as evident from the histogram, seem to be distributed in a similar way, obtain a very different contribu-

[3] The reason why there is such a large disagreement between A1 and A2 on BIR seems to be that A2, differently from A1, tends to always tag with BIR expressions relative to histological evaluation and expressions relative to benign diseases; this is likely an indication that the two annotators went through an insufficient alignment phase before starting the annotation. The high disagreement between the two annotators on the tag ECH can be instead due to the fact that it is a very infrequent tag, and it may be the case that too few instances of this tag were encountered in the alignment phase.

Table 3
Results of the multiple-annotator experiment: train the system on the union of `A1-only` and `A2-only`, and test it on the union of `Both(1)` and `Both(2)`. The first row reports the inter-annotator agreement values for A1 and A2 as measured on `Both`.

	BIR	ITE	IES	TFU	DEE	PAE	ECH	DEP	LLO	F_1^μ	F_1^M
A1 vs. A2	0.492	0.993	0.778	0.815	0.660	0.694	0.574	0.912	0.884	0.812	0.756
Baseline	0.537	0.994	0.768	0.866	0.710	0.688	0.618	0.775	0.863	0.821	0.758
Two-Stage	0.557	0.991	0.794	0.851	0.713	0.673	0.568	0.860	0.813	0.860	0.758
Ensemble	0.530	0.992	0.825	0.851	0.721	0.693	0.640	0.877	0.836	0.838	0.774
Baseline + PFs	0.516	0.983	0.794	0.873	0.698	0.679	0.645	0.843	0.861	0.828	0.766
Two-Stage + PFs	0.514	0.986	0.816	0.826	0.711	0.666	0.550	0.918	0.822	0.858	0.756
Ensemble + PFs	0.513	0.989	0.823	0.853	0.715	0.684	0.620	0.879	0.837	0.834	0.768

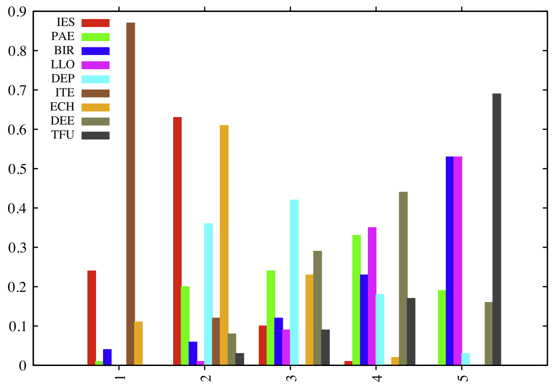


Fig. 3. Histogram representing, for each of the nine tags (indicated by nine different colors), the percentage of segments for the tag that occurs in a given quintile of the text. For example, the leftmost group of bars indicates the percentages of segments for each tag that occur in the first 20% of the text. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

tion from positional features, with the former improving from $F_1 = .772$ to $F_1 = .810$ and the latter deteriorating from $F_1 = .667$ to $F_1 = .614$.

Whether positional features are used or not, the baseline system and the two-stage system bring about similar accuracy in terms of F_1^M, but the two-stage system clearly outperforms the baseline in terms of F_1^μ (with a statistical significance level of $p < 0.1$). This indicates that the two-stage configuration performs better than the baseline at annotating frequent tags, and worse than the baseline at annotating rare tags, since microaveraged metrics are heavily influenced by the performance of the system on frequent tags (macroaveraged effectiveness measures pay instead equal attention to each tag independently of frequency). A clear example is the tag DEP, which is the rarest of all tags in our dataset (see Table 1), and on which the two-stage system performs radically worse than the baseline system.

Instead, the ensemble system obtains F_1^μ values close to the averages of the values obtained by the baseline and two-stage systems, and F_1^M values higher than those obtained by both systems. This indicates that the ensemble system is, on the whole, better than both its constituent systems, since it is better than the baseline when the tag is a frequent one, and it is much better than both systems on rare tags. In the configuration that does not use positional features, the ensemble system obtains $F_1^\mu = 0.858$ and $F_1^M = 0.793$, with an error reduction with respect to the baseline of 7.79% in terms of F_1^μ and of 6.75% on in terms of F_1^M. This improvement is confirmed to be significant by the statistical significance test with respect to both the baseline and the two-stage system (with $p < 0.005$ in both cases).

Concerning the difference among the performance obtained for the single tags, it is interesting to observe that the three tags for which the worst performance is obtained (BIR, ECH, and PAE) are also the ones (see Table 1) for which the segments have the smallest average length. (The same phenomenon will also be observed in the multiple-annotator experiments.) That long segments tend to be "easier" is explained by the fact that our evaluation measure is at the token level: since the difficult aspect in correctly tagging a segment is correctly spotting where the segment begins and ends, longer segments are evidently conducive to higher performance.

4.4.2. *Multiple-annotator results*

In the multiple-annotation scenario, the training set contains reports annotated by different annotators. Notwithstanding the alignment phase which the two annotators went through before starting their annotation work, each annotator may have a different annotation style, or a different understanding of the concepts which the tags represent. Therefore, similar segments may give rise to different tagging decisions by the two annotators. As a consequence, it is reasonable to assume that the resulting training set may possess a smaller internal consistency than in the single-annotator scenario; other things (e.g., number of training examples) being equal, this yields a more difficult task for the learning system.

The results of the multiple-annotation experiments are displayed in Table 3. The first observation we can make is that values are generally lower (the relative deterioration being mostly around 2%) than the corresponding values in Table 2. This indicates that, in the tradeoff between training data consistency and training data quantity mentioned in the previous section, the former seems to be more important: notwithstanding the fact that twice as many training examples are used in the multiple-annotator experiments, performance decreases, due to the higher level of internal inconsistency of the training set.

Concerning positional features, the multiple-annotator experiments confirm the results obtained in the single-annotator ones, since the use of positional features again does not bring about substantial differences. Significance tests confirms that the variations are not statistically significant. As in the single-annotator experiments, patterns are difficult to discern. Again, tags whose occurrence tends to be concentrated in a specific quantile of the texts do not seem to systematically benefit from the presence of positional features, and the very same tag (ECH) for which the positional features brought about a *deterioration* in F_1 from .667 to .614 in the single-annotation scenario, now even witnesses an *improvement* in F_1, from .618 to .645. The only conclusion one can draw is that more research needs to be done in order to assess how and whether positional features might benefit IE.

The improvement of the machine-learned systems with respect to the values of inter-coder agreement are smaller in magnitude than those observed in the single-annotator scenario, but are still substantial. This indicates that the learning algorithms are good at mediating between the different, sometimes contradictory information received from the annotators at training time, thus producing automatic annotations that mediate between the annotation styles of the individual annotators.

Concerning the comparison among the three learning methods, the results confirm those of the single-annotator scenario, with the two-stage method obtaining higher F_1^μ and lower F_1^M than the baseline, and the ensemble method mediating between the F_1^μ values obtained by the other two systems and obtaining the best F_1^M value. The ensemble method obtained an F_1^μ value of 0.838 and an F_1^M value of 0.774, thus bringing about an error reduction with respect to the baseline of 9.50% for F_1^μ and of 6.61% for F_1^M.

Here too, we have tested the statistical significance of the observed improvements using the same method discussed for the single-annotator experiments. In this case the test shows that the improvement of the two-stage method over the baseline is statistically significant with $p < 0.05$. The observed improvement of the ensemble method is statistically significant with respect both to the baseline ($p < 0.01$) and the two-stage method ($p < 0.05$).

4.5. A note on doing IE for resource-scarce languages

One of the main differences between the literature on information extraction from clinical texts (see Section 5) and our work, is that most of the former targets medical reports written in English, and as a consequence can leverage on the wide availability, for the English language, of lexical resources or "ontologies" specific to the domains of medicine or radiology. The present work may thus be taken as indicative of the level of extraction accuracy that can be attained for resource-scarce languages.

The availability of hierarchically organized domain-specific lexical resources is known to lead to higher accuracy in IE tasks. For instance, [9] reports obtaining a $F_1 = .800$ value on the i2b2 2010 challenge [7] by using CRFs with a standard set of features, and obtaining a $F_1 = .846$ value (thus reducing error by 23%) when features obtained by language-specific resources are brought into play. The reason is that, once such a resource is available, the vectorial representation of a token may be enhanced by using as features, additionally to the word the token is an instance of, its superordinate words in the hierarchy. This brings in higher statistical robustness (since the superordinate words, being more general, will tend to occur often), and reduces feature sparsity, which is beneficial to learning. Note that the same process of bringing in superordinate words cannot reliably be done via lexical resources that are not domain-specific, since non-technical terminology is more ambiguous, which may lead to selecting the *wrong* superordinates.

In experiments that we do not report here we have used "It(GT)-RadLex", a version of RadLex (a hierarchically organized controlled dictionary of radiology terms produced by the Radiological Society of North America[4]) that we have obtained by automatically translating RadLex into Italian via Google Translate. For each lexical expression (i.e., word or word n-gram) found in a report which exactly matched an entry in It(GT)-RadLex, the depth-k ancestor of the entry in It(GT)-RadLex (with depth 1 indicating the root) was added to the feature vectors representing the tokens in the lexical expression, for all $k = 2, 3, 4$. No improvement was observed in our experiments with respect to not using It(GT)-RadLex, which may indicate that more careful translations of the technical terms in a specialized dictionary are needed in order to make an impact on this IE task.

To the best of our knowledge no system for IE from radiology reports for the Italian language has been reported yet in the literature.

5. Related work

5.1. Information extraction

Gaizauskas et al. [18] give a good introduction to IE research carried out in the 1990s, including work spawned by the MUC series of evaluation challenges, while Sarawagi presents a more up-to-date survey [2]. McCallum [1] gives an excellent introduction to IE aimed at the non-specialist.

5.2. Information extraction from clinical text

The literature on IE from texts of medical interest can be roughly subdivided into (a) works about extracting information from the (bio-) medical literature (books, scientific articles, and the like) and (b) works about extracting information from texts generated during clinical activity (such as admission summaries, discharge summaries, clinical notes, and radiological reports). The former is an easier task than the latter, since clinical texts, unlike texts from the biomedical literature, are more informal, and thus abound in abbreviations (sometimes specific to the particular hospital or department where they originated), ungrammatical text, acronyms (even non-standard ones), and misspellings. Meystre et al. [19] review IE work specifically addressed to clinical narratives in the electronic health record,[5] while McNaught and Black [20] do the same for IE work addressed to biomedical literature. Note that the term "concept extraction" is often used in the biomedical area to actually refer to information extraction (see e.g., [12,21–23]); the latter is instead the standard name of the task, at least in the NLP community, where the task was first investigated.

Several works have been carried out in the field of information extraction from clinical narratives, and many of them use rule-based or dictionary-based systems in which the rules and dictionaries have been manually generated. Soderland et al. [24] extract instances of two concepts of medical interest, "diagnosis" and "symptoms", using a text analysis system and a dictionary induction system. Evans et al. [25] extract information about drug dosage, such as drug, dose level, and frequency, using a pattern-matching system. Harkema et al. [26] use an IE framework that selects relevant entities from a medical dictionary and then performs syntactic and semantic analyses over the text; the resulting information is then stored into a template that represents domain-specific entities and their properties and relations between them. Sotelsek-Margalef and Villena-Román [27] present a framework that uses IE from clinical reports to suggest medical diagnoses. Mykowiecka et al. [28] use a rule-based system that extracts information from clinical records written in Polish. Grishman et al. [29] use a rule-based system to detect disease outbreaks from clinical narratives.

Some works use a hybrid approach that combines the rule-based and the machine learning approaches. For example, Zhou et al. [30] extract three different types of information from semi-structured medical records, i.e., numerical values (e.g., age, blood pressure), medical terms (as from a patient's medical history), and categorical values (e.g., smoker/nonsmoker), for a total of twenty-four fields. They use two different approaches: for the first two types of information they use a combination of unsupervised pattern-matching methods, domain ontologies, and NLP techniques, while for the third type of information they use a machine learning method with input generated via NLP techniques. As another example, Taira et al. [31] use a hybrid rule-based/machine learning system to extract findings and their related properties using lexical resources and semantic parsing.

5.3. ML-based approaches to IE from clinical text

Only in recent years, also thanks to the interest generated by the i2b2 ("Informatics for Integrating Biology and the Bedside")

[4] http://rsna.org/RadLex.aspx.

[5] It has to be noted, however, that [19] takes a much wider view of what IE means, since it takes IE to also encompass other text-related tasks such as text classification.

challenges [7,8], researchers involved in IE from clinical narratives have started to massively use machine learning techniques. The first task of the 2011 i2b2/VA challenge [7] is indeed similar to the task we have faced in the present work, but is a bit simpler since (a) the concepts involved are neither nested nor potentially overlapping and (b) the narratives are in English, which is a resource-rich language.

There may be several reasons for the above-mentioned slow takeup of machine learning technology by the clinical NLP community. Torii et al. [12] conjecture that one of the reasons may be the fact that "phrases extracted from clinical text need to be normalized to fine-grained concepts, such as those defined in SNOMED CT and the Unified Medical Language." Certainly, one of the reasons is the fact that training data for clinical text has traditionally been scarce, often because of privacy and confidentiality issues; the datasets release within the context of the i2b2 challenge are probably the first publicly available datasets of their kind.

Most of the i2b2 participants that adopt a machine learning approach use CRFs of some sort. Torii et al. [12] use CRFs in a study of the portability of IE systems across multiple sources of clinical text; their CRFs system uses, aside from a standard set of text-derived features, features obtained via the use of the UMLS metathesaurus. Jiang et al. [9] use a CRFs system to perform named entity recognition as a preliminary step towards extracting concepts of clinical interest from discharge summaries. Jonnalagadda et al. [10] use a CRFs system in which the feature representation of a token is augmented with words that are "similar" to the word the token is an instance of, and where "similarity" is measured via distributional semantics. Patrick and Li [11] were the best performers in the i2b2 2009 challenge on extracting medication-related concepts from discharge summaries, by using CRFs to detect named entities and SVMs to further classify them. All of the above publications do not dwell on the details of the specific type of CRFs they actually use, which seems to indicate that they use the simple LC-CRFs type.

In terms of efforts not strictly related to i2b2, Li et al. [32] compare support vector machines (SVMs) and CRFs in the extraction of names of disorders from clinical text, and find that the latter are markedly superior to the former, obviously thanks to the fact that the former cannot encode dependencies among the labels of different tokens in the sequence. Wang and Patrick [33] use CRFs for named entity recognition in admission summaries; further application of a classifier committee formed by an SVMs classifier and a maximum entropy classifier, classifies the recognized named entities into classes.

5.4. Cascaded information extraction systems

Cascaded, multi-stage systems have been proposed in the past for several NLP tasks as applied to clinical texts. For instance, Jonnalagadda et al. [34] describes a multi-stage algorithm for coreference resolution (i.e., detecting if two linguistic expressions actually refer to the same entity) as applied to analysing medical discharge reports. In an attempt to automatically extract medication information from clinical records, Patrick and Li [11] use a CRFs learner and an SVMs learner arranged in a cascade, the CRFs learner being entrusted with recognizing named entities and the SVMs learner being entrusted with recognizing the relationship between two recognized entities. The above-mentioned work by Wang and Patrick [33] is in a similar vein. A trait in common among all these works is that they use CRFs in a first phase, usually as a recognition tool, and then pass the output of the CRFs to a tagger built via a different machine learning tool, usually as a classification tool. This is different from what we do, since our system revolves around the notion of using CRFs in both phases: a first phase where clauses are the units of interest, and a second phase in which such units are the individual tokens.

5.5. Positional features in information extraction

To the best of our knowledge, only another paper in the IE literature uses positional features: for a task of automatically extracting pros and cons of products from product reviews, Kim and Hovy [35] use positional features that indicate the first, the second, the last, and the second last sentence in a paragraph. Here the intuition is that, in a product review, pros and cons are important sentences that summarize the main point of the review. Our use is different, in the sense that we do not make any hypothesis of where important information for a given tag is located in a document, and simply record the relative position (as a percentage of the document length) in which the tagged segment is located. A fairly similar use of positional features is to be found in the work of Bramsen et al. [36]; however, the task addressed in [36] is not information extraction, but automatic segmentation of a document into temporally coherent segments.

6. Conclusions

We have presented novel solutions to the problem of extracting information from radiological reports (i.e., automatically annotating word sequences occurring in such reports according to a predefined set of concepts of interest) via supervised machine learning. Specifically, we have modified a standard linear-chain CRFs learning system in two novel ways, (i) cascading a clause-level LC-CRFs tagger and a token-level LC-CRFs tagger to obtain a two-stage system and (ii) organizing the resulting two-stage system and a traditional token-level LC-CRFs system into a confidence-weighted ensemble. We have also reported novel work in representing text by introducing "positional" features, i.e., features aiming to represent the quantiles of the text in which the instances of a given concept mainly tend to occur.

The single-annotator and multiple-annotator experiments we have conducted have shown that the positional features do not bring any substantial advantage. While the intuition that underlies them seems correct and promising, further experimentation will be needed to check if they can be of real use in some application context.

The same experiments have also shown (as also confirmed by statistical significance tests) that (i) the two-stage method is clearly superior to the baseline method for the most frequent tags, but is slightly inferior to it on the less frequent ones, and that (ii) the ensemble method is superior to the baseline on both frequent and infrequent tags. Interestingly, the accuracy levels obtained via these technologies are higher than the inter-coder agreement levels, as measured on the same test data and according to the same measure of agreement/accuracy. This is especially noteworthy since the feature set used in our work is fairly standard, and since it is widely believed that, as remarked by [12], "A tagger exploiting generic features can yield good performance, yet customized tokenization, features based on domain knowledge, hand-coded post-processing rules, and other fine-tuning can help improve performance". Our results, obtained via systems that use none of the above enhancements, show that machine-learned solutions nowadays reach effectiveness levels comparable to those of human experts. In particular, the experiments we have conducted do not use any specialized lexical resource or ontology, and are thus indicative of the level of accuracy that can be obtained in information extraction for resource-scarce languages.

There are several avenues for future research that this work leaves open. First of all, the findings presented in this paper would be strengthened if the same results could be replicated on other datasets, possibly made of mammography reports written in languages other than Italian, or possibly made of clinical narratives other than mammography reports. The current, limited public availability of datasets of clinical narratives is still an obstacle to progress in this direction, and initiatives like the i2b2 series of challenges already mentioned in Section 5 are a valiant step towards removing it. Second, the findings presented in Section 4.5 are an indication of the fact that the levels of accuracy obtainable when extracting information from documents written in resource-scarce languages are inferior to those obtainable for resource-rich languages such as English. Finding a way to bridge this gap, possibly leveraging recent results obtained in transfer learning/domain adaptation for cross-lingual text classification (see e.g., [37,38]), is an important research goal for future research.

Acknowledgments

This work has been funded by NoemaLife SpA in the framework of the ConnectToLife project. Thanks to the Istituto di Radiologia, Politecnico Umberto I, Roma, Italy, for making available the mammography report dataset used in this work; to Giulia Chiaruzzi and Cristiano Querzè for assistance in obtaining it; to Gianpiero Camilli, Michele Carenini, and Davide Distefano, for encouraging this work; and to Emanuele Pianta, Christian Girardi and Roberto Zanoli for giving us access to their TextPro system. This work is dedicated to the fond memory of Emanuele Pianta, a valuable colleague and dear friend who left us far too early.

Appendix A. A baseline CRFs system for IE

LC-CRFs are members of the class of *graphical models*, a family of probability distributions that factorize according to an underlying graph [39]. Given a text $\mathbf{x}$, LC-CRFs model the conditional probability $p(\mathbf{y}|\mathbf{x})$ of a vector $\mathbf{y}$ of labels given $\mathbf{x}$ according to the formula

$$p(\mathbf{y}|\mathbf{x}) = \frac{1}{Z(\mathbf{x})}\prod_{t=1}^{|\mathbf{x}|}\Psi_u(y_t,\mathbf{x}_t)\prod_{t=1}^{|\mathbf{x}|-1}\Psi_b(y_t,y_{t+1},\mathbf{x}_t) \tag{A.1}$$

where $\frac{1}{Z(\mathbf{x})}$ is a normalization factor which ensures that the final result is indeed a probability and Ψ_u are the *unigram factors* of the feature vectors $\mathbf{x}_t$ and the labels y_t, and Ψ_b are the *bigram factors* of the feature vectors $\mathbf{x}_t$ and the labels y_t and y_{t+1}.

Unigram and bigram factors are defined, respectively, as

$$\Psi_u(y_t,\mathbf{x}_t) = \exp\sum_k \theta_k f_k(y_t,\mathbf{x}_t) \tag{A.2}$$

$$\Psi_b(y_t,y_{t+1},\mathbf{x}_t) = \exp\sum_h \theta_h f_h(y_t,y_{t+1}\mathbf{x}_t) \tag{A.3}$$

where f_k and f_h are binary *feature functions*, i.e., functions that return 0 or 1 depending on the values of their arguments (see Eqs. (A.4) and (A.5) for a more detailed explanation) and the θ_k's and θ_h's (one for each feature function) are numerical parameters to be estimated from the training data.

The feature functions are binary functions that describe dependencies between features and variables as detected in the training data. In this work we have used the following two types of feature function:

$$f_{idv}(y_t,\mathbf{x}_t) = \begin{cases} 1 & \text{if } y_t = i \text{ and } \mathbf{x}_{td} = v \\ 0 & \text{otherwise} \end{cases} \tag{A.4}$$

$$f_{ijdv}(y_t,y_{t+1},\mathbf{x}_t) = \begin{cases} 1 & \text{if } y_t = i \text{ and } y_{t+1} = j \text{ and } \mathbf{x}_{td} = v \\ 0 & \text{otherwise} \end{cases} \tag{A.5}$$

where there is a function of type $f_{idv}(y_t,\mathbf{x}_t)$ and a function of type $f_{ijdv}(y_t,y_{t+1},\mathbf{x}_t)$ for each combination of $i \in \{c,\bar{c}\}$, $j \in \{c,\bar{c}\}$, $d \in \{1,\ldots,\Omega\}$, and actual value v that d takes in the training data. For simplicity of notation, in Eqs. (A.2) and (A.3) we have used indexes k and h to range on all the possible values of the triple (i,d,v) and of the quadruple (i,j,d,v), respectively.

The following is an example of a feature function of the type described in Eq. (A.4):

$$f_{PAE,word,\texttt{formazioni}}(y_t,\mathbf{x}_t) = \begin{cases} 1 & \text{if } y_t = PAE \\ & \text{and } \mathbf{x}_{t,word} = \texttt{formazioni} \\ 0 & \text{otherwise} \end{cases} \tag{A.6}$$

where *PAE* is a tag, *word* is a dimension in the feature vector, and `formazioni` an actual value that occurs for dimension *word* in the training data. It can be paraphrased as saying "If the label of the current token is *PAE*, and the current token is an occurrence of the word `formazioni`, then return 1, else return 0". An example feature function of the type described in Eq. (A.5) is instead

$$f_{PAE,PAE,POS,adjective}(y_t,y_{t+1},\mathbf{x}_t) = \begin{cases} 1 & \text{if } y_t = PAE \\ & \text{and } y_{t+1} = PAE \\ & \text{and } \mathbf{x}_{t,POS} = \text{adjective} \\ 0 & \text{otherwise} \end{cases} \tag{A.7}$$

with the obvious meaning. The experimenter can decide which properties of the text to take into account (e.g., capitalized words, parts of speech, prefixes, suffixes, and so on) by including feature functions in which these properties play the role of dimension d in Eqs. (A.4) and (A.5).

Functions of the type formalized by Eq. (A.4) describe the relation that occurs between the label of a token and the value of the dth feature of the vector representing it. Functions of the type formalized by Eq. (A.5) describe instead the relation that occurs between the label of a token, the label of the token that immediately follows it, and the dth feature of the vector representing the former.

In both types of feature functions the clause "and $\mathbf{x}_{td} = v$" may be missing. With the feature functions of Eq. (A.4) it is thus possible to model the relative frequency of a given label, while with the feature functions of Eq. (A.5) it is thus possible to model the relative frequency of a given sequence of two consecutive labels.

It is important to notice that all the information that is needed for computing the feature functions of Eqs. (A.4) and (A.5) must be contained in the feature vector $\mathbf{x}_t$ that represents token x_t. In particular, this vector may contain information about the token at position t, but also information about the token that precedes (or follows) it.

Fig. 4 is thus the "factor graph" that represents LC-CRFs, where circles represent the variables of the distribution and squares represent the factors into which the distribution is decomposed.

In LC-CRFs the learning phase consists in estimating the parameters $\Theta = \{\{\theta_k\}_k \cup \{\theta_h\}_h\}$ from a training set $Tr = \{\mathbf{x}^1,\ldots,\mathbf{x}^{|Tr|}\}$ of labelled texts, with text $\mathbf{x}^s$ labelled by vector $\mathbf{y}^s$. There is one parameter θ_k for each different combination of $i \in \{c,\bar{c}\}$, $d \in [1,\ldots,\Omega]$, and actual values v that the dth dimension of the feature vector takes in the training set (similarly for θ_h). Parameter estimation is achieved by maximizing a regularized version of the conditional log likelihood of $p(\mathbf{y}|\mathbf{x})$, which, given Eq. (A.1), amounts to:

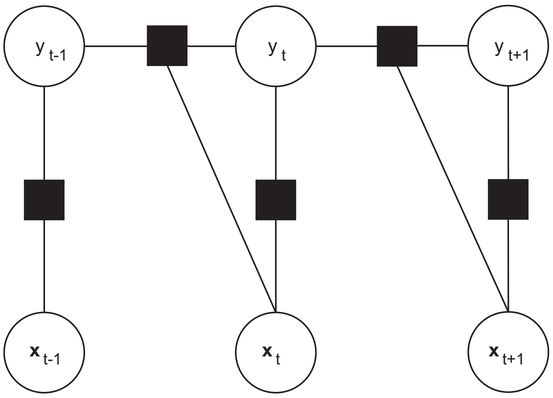


Fig. 4. Factor graph of LC-CRFs.

$$
\begin{aligned}
\ell(\theta) &= \sum_{s=1}^{|Tr|} \log p(\mathbf{y}^s|\mathbf{x}^s) - \sum_{\theta\in\Theta}\frac{\theta^2}{2\delta^2} \\
&= \sum_{s=1}^{|Tr|}\sum_{t=1}^{|\mathbf{x}^s|}\sum_{k}\theta_k f_k(y_t^s,\mathbf{x}_t^s) + \sum_{t=1}^{|\mathbf{x}^s|-1}\sum_{h}\theta_h f_h(y_t^s,y_{t+1}^s,\mathbf{x}_t^s) \\
&\quad - \sum_{s=1}^{|Tr|}\log Z(\mathbf{x}^s) - \sum_{\theta\in\Theta}\frac{\theta^2}{2\delta^2} \qquad \text{(A.8)}
\end{aligned}
$$

where δ is a regularization parameter that determines how much θ is penalized in order to avoid overfitting.

Maximizing the regularized conditional log likelihood of $p(\mathbf{y}|\mathbf{x})$ is an optimization problem which in LC-CRFs is typically solved by means of numerical methods such as gradient ascent, Newton and quasi-Newton methods, and iterative scaling. Once the parameters Θ that identify the model have been estimated, a new text $\mathbf{x}$ is labelled by applying the Viterbi algorithm, which predicts the most probable label sequence $\mathbf{y}^*$ given a model and a sequence $\mathbf{x}$, i.e.,

$$\mathbf{y}^* = \arg\max_{\mathbf{y}} p(\mathbf{y}|\mathbf{x}) \qquad \text{(A.9)}$$

where $p(\mathbf{y}|\mathbf{x})$ is as in Eq. (A.1). That is, of all possible vectors $\mathbf{y}$ consisting of $|\mathbf{x}|$ labels drawn from $\{c,\bar{c}\}$, the vector $\mathbf{y}^*$ with the maximum probability conditioned on $\mathbf{x}$ is chosen.

References

[1] McCallum A. Information extraction: distilling structured data from unstructured text. Queue 2005;3(9):48–57.

[2] Sarawagi S. Information extraction. Found Trends Databases 2008;1(3):261–377.

[3] Uzuner Ö, Luo Y, Szolovits P. Evaluating the state of the art in automatic de-identification. J Am Med Inform Assoc 2007;14(5):550–63.

[4] Lafferty J, McCallum A, Pereira F. Conditional random fields: probabilistic models for segmenting and labeling sequence data. In: Proceedings of the 18th international conference on machine learning (ICML 2001), Williamstown, USA; 2001. p. 282–9.

[5] Sutton C, McCallum A. An introduction to conditional random fields for relational learning. In: Getoor L, Taskar B, editors. Introduction to statistical relational learning. Cambridge (USA): The MIT Press; 2007. p. 93–127.

[6] Sutton C, McCallum A. An introduction to conditional random fields. Found Trends Mach Learn 2012;4(4):267–373.

[7] Uzuner Ö, South BR, Shen S, DuVall SL. 2010 i2b2/VA challenge on concepts assertions and relations in clinical text. J Am Med Inform Assoc 2011;18(5):552–6.

[8] Uzuner Ö, Solti I, Cadag E. Extracting medication information from clinical text. J Am Med Inform Assoc 2010;17(5):514–8.

[9] Jiang M, Chen Y, Liu M, Rosenbloom ST, Mani S, Denny JC, et al. A study of machine-learning-based approaches to extract clinical entities and their assertions from discharge summaries. J Am Med Inform Assoc 2011;18(5):601–6.

[10] Jonnalagadda S, Cohen T, Wu S, Gonzalez G. Enhancing clinical concept extraction with distributional semantics. J Biomed Inform 2012;45(1):129–40.

[11] Patrick J, Li M. High accuracy information extraction of medication information from clinical notes: 2009 i2b2 medication extraction challenge. J Am Med Inform Assoc 2010;17:524–7.

[12] Torii M, Wagholikar K, Liu H. Using machine learning for concept extraction on clinical documents from multiple data sources. J Am Med Inform Assoc 2011;18(5):580–7.

[13] Roli F, Giacinto G, Vernazza G. Methods for designing multiple classifier systems. In: Proceedings of the 2nd international workshop on multiple classifier systems (MCS 2001), Cambridge, UK; 2001. p. 78–87.

[14] Esuli A, Sebastiani F. Evaluating information extraction. In: Proceedings of the conference on multilingual and multimodal information access evaluation (CLEF 2010), Padova, Italy; 2010. p. 100–11.

[15] Suzuki J, McDermott E, Isozaki H. Training conditional random fields with multivariate evaluation measures. In: Proceedings of the 21st international conference on computational linguistics and 44th annual meeting of the ACL (ACL/COLING 2006), Sydney, Australia; 2006. p. 217—24.

[16] Cunningham H. GATE a general architecture for text engineering. Comput Human 2002;36(2):223–54.

[17] Pianta E, Girardi C, Zanoli R. The TextPro tool suite. In: Proceedings of the 6th language resources and evaluation conference (LREC 2008), Marrakech, Morocco; 2008.

[18] Gaizauskas R, Wilks Y. Information extraction: beyond document retrieval. J Document 1998;54(1):70–105.

[19] Meystre SM, Savova GK, Kipper-Schuler KC, Hurdle JF. Extracting information from textual documents in the electronic health record: a review of recent research. In: Geissbuhler A, Kulikowski C, editors. IMIA yearbook of medical informatics. Stuttgart (DE): Schattauer Publishers; 2008. p. 128–44.

[20] McNaught J, Black W. Information extraction: the task. In: Ananiadou S, McNaught J, editors. Text mining for biology and biomedicine. London (UK): Artech House Books; 2006. p. 143–76.

[21] Bleik S, Xiong W, Wang Y, Song M. Biomedical concept extraction using concept graphs and ontology-based mapping. In: Proceedings of the 4th IEEE international conference on bioinformatics and biomedicine (BIBM 2010), Hong Kong, China; 2010. p. 553–6.

[22] Dinh D, Tamine L. Biomedical concept extraction based on combining the content-based and word order similarities. In: Proceedings of the 26th ACM symposium on applied computing, TaiChung, Taiwan; 2011. p. 1159–63.

[23] Kang N, Afzal Z, Singh B, van Mulligen EM, Kors JA. Using an ensemble system to improve concept extraction from clinical records. J Biomed Inform 2012;45(3):423–8.

[24] Soderland S, Aronow D, Fisher D, Aseltine J, Lehnert W. Machine learning of text analysis rules for clinical records. Tech rep. TE-39. Amherst (USA): Center for Intelligent Information Retrieval, University of Massachusetts; 1995.

[25] Evans DA, Brownlow ND, Hersh WR, Campbell EM. Automating concept identification in the electronic medical record: an experiment in extracting dosage information. In: Proceedings of the annual fall symposium of the American Medical Informatics Association, Washington, USA; 1996. p. 388–92.

[26] Harkema H, Roberts I, Gaizauskas R, Hepple M. Information extraction from clinical records. In: Proceedings of the 4th UK e-science all hands meeting (AHM 2005), Nottingham, UK; 2005. p. 39–43.

[27] Sotelsek-Margalef A, Villena-Román J. MIDAS: an information-extraction approach to medical text classification. Proc Lenguaje Nat 2008;41:97–104.

[28] Mykowiecka A, Marciniak M, Kupść A. Rule-based information extraction from patients' clinical data. J Biomed Inform 2009;42(5):923–36.

[29] Grishman R, Huttunen S, Yangarber R. Information extraction for enhanced access to disease outbreak reports. J Biomed Inform 2002;35(4):236–46.

[30] Zhou X, Han H, Chankai I, Prestrud AA, Brooks AD. Converting semi-structured clinical medical records into information and knowledge. In: Proceedings of the 21st international conference on data engineering (ICDE 2005), Tokyo, Japan; 2005. p. 1162–9.

[31] Taira RK, Soderland SG, Jakobovits RM. Automatic structuring of radiology free-text reports. RadioGraphics 2001;21(1):237–45.

[32] Li D, Kipper-Schuler K, Savova G. Conditional random fields and support vector machines for disorder named entity recognition in clinical texts. In: Proceedings of the ACL workshop on current trends in biomedical natural language processing (BioNLP 2008), Columbus, USA; 2008. p. 94–5.

[33] Wang Y, Patrick J. Cascading classifiers for named entity recognition in clinical notes. In: Proceedings of the RANLP 2009 workshop on biomedical information extraction, Borovets, Bulgaria; 2009. p. 42–9.

[34] Jonnalagadda SR, Li D, Sohn S, Wu ST, Wagholikar K, Torii M, et al. Coreference analysis in clinical notes: a multi-pass sieve with alternate anaphora resolution modules. J Am Med Inform Assoc 2012;19(5):867–74.

[35] Kim S-M, Hovy E. Automatic identification of pro and con reasons in online reviews. In: Proceedings of the 21st international conference on computational linguistics and 44th annual meeting of the Association for Computational Linguistics (COLING/ACL 2006), Sydney, Australia; 2006. p. 483–90.

[36] Bramsen P, Deshpande P, Lee YK, Barzilay R. Finding temporal order in discharge summaries. In: Proceedings of the 30th AMIA annual symposium (AMIA 2006), Washington, USA; 2006. p. 81–5.

[37] Pan W, Zhong E, Yang Q. Transfer learning for text mining. In: Aggarwal CC, Zhai C, editors. Mining text data. Heidelberg (DE): Springer; 2012. p. 223–58.

[38] Wang H, Huang H, Nie F, Ding CH. Cross-language web page classification via dual knowledge transfer using nonnegative matrix tri-factorization. In: Proceedings of the 34th ACM international conference on research and development in information retrieval (SIGIR 2011), Beijing, China; 2011. p. 933–42.

[39] Wainwright MJ, Jordan MI. Graphical models, exponential families, and variational inference. Found Trends Mach Learn 2008;1(1/2):1–305.