Evaluating the state of the art in coreference resolution for electronic medical records

Ozlem Uzuner,[1] Andreea Bodnari,[2] Shuying Shen,[3,4,5] Tyler Forbush,[3,4] John Pestian,[6] Brett R South[3,4,5]

▸ Additional materials are published online only. To view these files please visit the journal online (http://dx.doi.org/10.1136/amiajnl-2011-000784).

[1]Department of Information Studies, University at Albany, SUNY, Albany, New York, USA
[2]Computer Science and Artificial Intelligence Laboratory, MIT, Cambridge, Massachusetts, USA
[3]VA Salt Lake City Health Care System, Salt Lake City, Utah, USA
[4]Department of Internal Medicine, University of Utah, Salt Lake City, Utah, USA
[5]Department of Biomedical Informatics, University of Utah, Salt Lake City, Utah, USA
[6]Cincinnati Children's Hospital, University of Cincinnati Computational Medicine Center, Cincinnati, Ohio, USA

Correspondence to
Dr Ozlem Uzuner, Department of Information Studies, College of Computing and Information, University at Albany, SUNY, Draper 114A, 135 Western Avenue, Albany, NY 12222, USA; ouzuner@albany.edu

Received 18 December 2011
Accepted 23 January 2012
Published Online First
24 February 2012

ABSTRACT

Background The fifth i2b2/VA Workshop on Natural Language Processing Challenges for Clinical Records conducted a systematic review on resolution of noun phrase coreference in medical records. Informatics for Integrating Biology and the Bedside (i2b2) and the Veterans Affair (VA) Consortium for Healthcare Informatics Research (CHIR) partnered to organize the coreference challenge. They provided the research community with two corpora of medical records for the development and evaluation of the coreference resolution systems. These corpora contained various record types (ie, discharge summaries, pathology reports) from multiple institutions.

Methods The coreference challenge provided the community with two annotated ground truth corpora and evaluated systems on coreference resolution in two ways: first, it evaluated systems for their ability to identify mentions of concepts and to link together those mentions. Second, it evaluated the ability of the systems to link together ground truth mentions that refer to the same entity. Twenty teams representing 29 organizations and nine countries participated in the coreference challenge.

Results The teams' system submissions showed that machine-learning and rule-based approaches worked best when augmented with external knowledge sources and coreference clues extracted from document structure. The systems performed better in coreference resolution when provided with ground truth mentions. Overall, the systems struggled in solving coreference resolution for cases that required domain knowledge.

INTRODUCTION

The fifth i2b2/VA Workshop on Natural Language Processing Challenges for Clinical Records, organized by Informatics for Integrating Biology and the Bedside (i2b2) and the Veterans Affairs (VA) Consortium for Healthcare Informatics Research (CHIR), gathered the natural language processing (NLP) community for resolving coreference in electronic medical records. We refer to this challenge as the coreference resolution challenge. This article presents an overview of the coreference resolution challenge, data, and evaluation metrics; it reviews and evaluates the systems developed for this challenge, and provides directions for future research in clinical coreference resolution.

Coreference resolution determines whether two concepts are coreferent, that is, linked by an 'identity' or 'equivalence' relation. For example, in the sentence *'She was scheduled to receive a temporal artery biopsy, but she never followed up on that testing,'* 'a temporal artery biopsy' and 'that testing' are equivalent because they refer to the same entity. We refer to the two textual occurrences of the concepts 'a temporal artery biopsy' and 'that testing' as mentions; two or more equivalent mentions create a coreference chain. We refer to mentions that are not involved in any coreference chains as singletons. The goal of the coreference resolution challenge was to encourage the development of systems that could identify coreference chains. For this purpose, i2b2/VA provided coreference annotation guidelines (see online supplementary appendix I), concept mention annotations, and a training set of ground truth coreference chains. Twenty teams from 29 organizations and nine countries participated in the coreference resolution challenge (see online supplementary table 1). The results of the challenge were presented in a workshop that i2b2/VA organized with the Computational Medicine Center of Cincinnati Children's Hospital, in co-sponsorship with the American Medical Informatics Association (AMIA), at the Fall Symposium of AMIA in 2011.

RELATED WORK

In the NLP literature, coreference resolution has focused primarily on the newspaper[1 2] and biomedical corpora,[3] leaving the clinical corpora relatively unexplored (see online supplements of Bodnari *et al*[4] for related work in the open domain).[5–7] The work of He[8] explored coreference resolution in discharge summaries using a supervised decision-tree classifier and a carefully selected set of features. Zheng *et al*[7] carried out a comprehensive review of coreference resolution methodologies in the open domain and suggested transferring these techniques to the clinical domain.

The coreference resolution challenge continued i2b2's efforts to release annotated clinical records to the NLP community for the advancement of the state of the art. This challenge built on past i2b2 challenges,[9–13] as well as past NLP shared-task efforts outside the clinical domain.[14] The first i2b2 challenge proposed an information extraction task targeting the de-identification of protected health information[9] and a document classification task targeting the smoking status of patients.[10] The second i2b2 challenge proposed a multi-document classification task focused on obesity and 15 of its co-morbidities. This challenge encompassed several NLP tasks: information extraction for disease-specific details, negations and uncertainty extraction on diseases, and classification of patient records.[11] The third i2b2 challenge targeted the extraction of medication and medication-related information.[12] The fourth i2b2 challenge proposed three tasks: a concept extraction task targeting the extraction of medical concepts from clinical records; an assertion classification task targeting the assignment of assertion types for medical problem concepts; and a relation classification task targeting the assignment of relation types that hold between medical concepts. The fifth i2b2/VA challenge (ie, coreference resolution challenge) extends relation classification to coreference resolution.

DATA

The data for the coreference challenge consisted of two separate corpora: the i2b2/VA corpus and the Ontology Development and Information Extraction (ODIE) corpus. The i2b2/VA corpus contained de-identified discharge summaries from Beth Israel Deaconess Medical Center, Partners Healthcare, and University of Pittsburgh Medical Center (UPMC). In addition, UPMC contributed de-identified progress notes to the i2b2/VA corpus. The ODIE corpus contained de-identified clinical reports and pathology reports from Mayo Clinic, and de-identified discharge records, radiology reports, surgical pathology reports, and other reports from UPMC. Table 2 shows the number of reports from each institution and the division of reports into training and test sets in these corpora. The i2b2/VA corpus was produced by i2b2 and the VA, and the ODIE corpus was produced under the ODIE grant and was donated to the i2b2/VA challenge under SHARP—Secondary Use of Clinical Data from the Office of the National Coordinator (ONC) for Health Information Technology.[15]

The ODIE corpus contained 10 concept categories: anatomical site, disease or syndrome, indicator/reagent/diagnostic aid, laboratory or test result, none, organ or tissue function, other, people, procedure, and sign or symptom.[16] In comparison, the i2b2/VA corpus contained five concept categories: problem, person, pronoun, test, and treatment.[13] For annotation details on the ODIE corpus refer to Savova *et al*.[15]

Each record in the i2b2/VA corpus was annotated by two independent annotators for coreference pairs. Then the pairs were post-processed in order to create coreference chains. These chains were presented to an adjudicator, who resolved the disagreements between the original annotations, and added or deleted annotations as necessary. The outputs of the adjudicators were then re-adjudicated, with particular attention being paid to duplicates and enforcing consistency in the annotations.

Table 2 ODIE and i2b2/VA train and test file counts

Corpus	File source	Training files	Test files	Total
ODIE	Mayo Clinic	58	39	97
	Clinical reports	30	19	49
	Pathology reports	28	20	48
	University of Pittsburgh Medical Center	40	27	67
	Discharge reports	10	6	16
	Radiology reports	11	7	18
	Surgical pathology reports	10	8	18
	Other reports	9	6	15
	Total	98	66	164
i2b2/VA	Beth Israel Deaconess Medical Center	115	79	194
	Partners Healthcare	136	94	230
	University of Pittsburgh Medical Center	241	149	390
	Discharge summaries	119	77	196
	Progress notes	122	72	194
	i2b2/VA without Pittsburgh	251	173	424
	Total	492	322	814
ODIE and i2b2/VA	Total	590	388	978

Appendix II and table 3 in the online supplements contain annotation details and inter-annotator agreement results for the i2b2/VA corpus.

The ODIE corpus contained 419 chains, with an average chain length of 5.671 concept mentions and maximum chain length of 90 mentions (see table 4). The i2b2/VA corpus contained 5227 chains, with an average chain length of 4.326 concept mentions and maximum chain length of 122 concept mentions (see table 4).

Both the ODIE and i2b2/VA corpora were released under a data use agreement that allows their use for research beyond the coreference challenge. The data use agreement is available at https://www.i2b2.org/NLP/Coreference/Agreement.php. All relevant institutional review boards approved this challenge and the use of the de-identified clinical records.

METHODS

The coreference challenge consisted of three tasks. The first task (Task 1A) focused on mention extraction and coreference resolution on the ODIE corpus. The systems participating in this task had to first identify mentions from raw text and then perform coreference resolution on these mentions. The second task (Task 1B) focused on coreference resolution on the ODIE corpus using ground truth concept mentions and the raw text of the ODIE clinical records. The third task (Task 1C) focused on coreference resolution on the i2b2/VA corpus using the ground truth concept mentions and the raw text of the i2b2/VA clinical records.

For Task 1C, four out of the 20 participating teams could not obtain UPMC records. We consequently ran two separate evaluations for Task 1C. The first evaluation was run on the entire i2b2/VA corpus (Task 1C i2b2) and included only the 16 teams who could obtain all of the i2b2/VA data. The second evaluation was run on the i2b2/VA corpus without the UPMC records and included all 20 teams (Task 1C i2b2/UPMC).

Each team could submit up to three system outputs per task and was evaluated on their best performing output per task.

Evaluation metrics for mention extraction

Following the evaluation methodology of the fourth i2b2/VA challenge,[13] we evaluated the systems' performance on mention extraction using precision, recall, and F-measure. We considered

Table 4 ODIE and i2b2/VA chain count, chain average length, and chain maximum length

Corpus	File source	Chain count	Chain average length	Chain maximum length
ODIE	Mayo Clinic	208	5.519	72
	Clinical reports	147	6.510	72
	Pathology reports	61	3.131	7
	University of Pittsburgh Medical Center	211	5.820	90
	Discharge records	55	6.091	58
	Other reports	92	7.000	90
	Radiology reports	26	3.615	9
	Surgical pathology reports	38	4.079	18
	ODIE	419	5.671	90
i2b2/VA	Beth Israel Deaconess Medical Center	1816	4.155	122
	Partners Healthcare	1395	4.352	105
	University of Pittsburgh Medical Center	2016	4.461	121
	Discharge summaries	1103	4.509	104
	Progress notes	913	4.403	121
	i2b2/VA without Pittsburgh	3211	4.241	122
	i2b2/VA	5227	4.326	122

both exact and at least partial mention overlap with the ground truth mentions (see online supplements for details). Evaluation of mention extraction was performed for Task 1A only.

Evaluation metrics for coreference resolution

We evaluated the systems' performance on coreference resolution using three evaluation metrics: MUC,[17] B-CUBED,[18] and CEAF.[19] Each metric presents different strengths and weaknesses. We used the unweighted average of the MUC, B-CUBED, and CEAF metrics as a measure of coreference performance on chains. We evaluated systems across all semantic categories at the same time, without a distinction in semantic category. For Task 1A, we gave systems credit for only the pairs and chains that contained mentions that matched the ground truth exactly, that is, exact overlap. We then repeated the same evaluation for mentions with at least partial overlap. For Task 1B and 1C, we performed coreference evaluation of the system chains against the ground truth.

MUC metrics

MUC metrics evaluated the set of system chains by looking at the minimum number of pair additions and removals required for them to match the ground truth chains.[17] The pairs to be added represented false negatives, while the pairs to be removed represented false positives. Let K represent the ground truth chains set, and R the system chains set. Given chains k and r from K and R, respectively, MUC recall and precision of R were:

$$recall = \frac{\sum_{k}(|k| - m(k,R))}{\sum_{k}(|k| - 1)}$$

$$precision = \frac{\sum_{k}(|r| - m(k,K))}{\sum_{k}(|r| - 1)}$$

$m(r,K)$, by definition, represented the number of chains in K that intersected the chain r.

The MUC F-measure was given by:

$$F - measure = \frac{2*recall*precision}{recall + precision}$$

B-CUBED metrics

B-CUBED metrics evaluated system performance by measuring the overlap between the chains predicted by the system and the ground truth chains.[18] Let C be a collection of N documents, d a document in C, and m a markable in document d. We defined the ground truth chain that included m as G_m and the system chain that contained m as S_m. O_m was the intersection of G_m and S_m. B-CUBED recall and precision were defined as:

$$recall = \frac{1}{N}\sum_{d\in C}\sum_{m\in d}\frac{|O_m|}{|G_m|}$$

$$precision = \frac{1}{N}\sum_{d\in C}\sum_{m\in d}\frac{|O_m|}{|S_m|}$$

The B-CUBED F-measure was identical to the MUC F-measure.

CEAF metrics

CEAF metrics first computed an optimal alignment ($\Phi(g^*)$) between the system chains and the ground truth chains based on a similarity score. This score could be based on the mentions or on the chains. The chains-based score had two variants, ϕ_3 and ϕ_4; we employed ϕ_4, unless otherwise specified.[19]

Let gold standard chains in a document d be $K(d)=\{K_i: i=1,2\ldots,\{K(d)\}\}$, and system chains in a document d be $R(d)=\{R_i{:}i=1,2\ldots,|R(d)|\}$. Let K_i and R_i be chains in $K(d)$ and $R(d)$, respectively. The chain-based scores were defined as:

$$\phi_3(K_i,R_j) = |K_i \cap R_j|$$

$$\phi_4\left(K_i,R_j\right) = \frac{2|K_i \cap R_j|}{|K_i| + |R_j|}$$

The CEAF precision and recall were defined as:

$$precision = \frac{\Phi(g^*)}{\sum_i \phi(R_i,R_i)}$$

$$recall = \frac{\Phi(g^*)}{\sum_i \phi(K_i,K_i)}$$

The CEAF F-measure was identical to the MUC F-measure.

Significance tests

We assessed whether two system outputs were significantly different from each other by using approximate randomization tests.[20] Let A and B be two different systems, with outputs of j chains and k chains, respectively. We evaluated systems A and B using the unweighted average F-measure and computed the absolute difference between the unweighted average F-measure of system A (f_A) and unweighted average F-measure of system B (f_B) as $f=|f_A-f_B|$. We collected the chains of system A and the chains of system B; we created a superset C, of M=j+k chains. We then performed step 1 and step 2 N times, as described below. In step 1, we selected from C j chains randomly and without resampling and created the pseudoset of chains A_p. The remaining k chains in C created the pseudoset of chains B_p. In step 2, we computed the absolute difference of f_A', the unweighted average F-measure of A_p, and f_B', the unweighted average F-measure of B_p, as $f_p = |f_A' - f_B'|$. At the end of the N iterations, we computed Nt, the number of times that $|f_p-f|>=0$ and calculated the p value between A and B as p=(Nt+1)/(N+1). We ran significance tests with N=100 and α=0.01.

SYSTEMS

The 2011 i2b2/VA challenge systems were grouped with respect to their use of external resources, involvement of medical experts, and methods (see online supplements for definitions). Seven systems were described by their authors as rule-based, eight systems as supervised, and three as hybrid. Two systems were declared to have utilized external resources, and two systems were designed under the supervision of medical experts.

In general, the 2011 i2b2/VA challenge systems created separate modules to solve coreference for the person concepts, pronoun concepts, and the non-person concepts (ie, problem, test, treatment, etc). To aid coreference resolution for the person category, most systems distinguished between the patient and non-patient entities. All systems explored the context surrounding the mentions. Below we provide more details for the rule-based, hybrid, and supervised systems developed for the i2b2/VA challenge.

Rule-based 2011 i2b2/VA challenge systems

In general, the rule-based systems assumed that two mentions were more likely to corefer if located in the same document section. Then, these systems used regular expressions, handcrafted keywords, and internet searches to classify mentions into patient, medical personnel, and family member groups. The personal pronouns were assumed to corefer to the closest person mentions, while the non-personal pronouns were classified based on their form and syntactic dependency relations. To resolve coreference in non-person categories, the systems used token overlap; some also incorporated external knowledge. Gooch[21] and Grouin *et al*[22] integrated semantic disambiguation, spelling correction, and abbreviation extension derived from Wikipedia abbreviations. Hinote *et al*[23] used semantic clues like dates, locations, and descriptive modifiers. They also used Wordnet[24] synonyms to match words within the mentions, the UMLS database[25] to determine closely related medical mentions, and automatic internet searches to determine whether a mention referred to medical personnel. Yang *et al*[26] incorporated a preprocessing step into their system; they parsed, tagged, and normalized the raw texts before coreference resolution.

Hybrid 2011 i2b2/VA challenge systems

The coreference challenge had three hybrid systems: two multisieve classifiers (Jonnalagadda *et al*[27] and Rink *et al*[28]) and a pairwise classifier (Jindal *et al*[29]). Jonnalagadda *et al* experimented on pronoun classification using a rule-based and a factorial hidden Markov model classifier. Rink *et al* adjusted the first pass of their multi-sieve model to identify the patient mentions. These mentions were then combined into a single coreference chain. Jindal *et al* classified mention pairs containing one pronoun separately from mention pairs containing two pronouns; they also differentiated between the patient, doctor, and family member instances of the person category. The hybrid systems resolved coreference for the non-person mentions by incorporating external domain knowledge. Rink *et al* employed Wikipedia aliases for marking alternative spellings and identifying synonyms. Jonnalagadda *et al* and Jindal *et al* extracted abbreviations, synonyms, and other relations from UMLS. Jindal *et al* used system features such as anatomical terms corresponding to body location and body parts.

Supervised 2011 i2b2/VA challenge systems

Much like the rule-based and hybrid systems, the supervised coreference resolution systems paid special attention to the person and pronoun categories. In general, these systems tried to distinguish the patient mentions from other person mentions. Coreference resolution for non-person categories used features like mention similarity and document section. Anick *et al*[30] applied these features with a maximum entropy classifier and added time frame and negation. Cai *et al*[31] applied a weakly supervised, graph-based model. Xu *et al*[32] chose a support vector machine (SVM) classifier enhanced with features from the document's structure and world knowledge from sources like Wikipedia,[33] Probase,[34] and NeedleSeek.[35] Xu *et al* used a mapping engine with additional features like anatomy and position, medication information, time, and space.

RESULTS

Three systems participated in Task 1A, eight systems participated in Task 1B, and 20 systems participated in Task 1C. All systems were evaluated on held out test data for their task. In order to analyze systems' performance against a reference standard, we defined a baseline system that predicted all mentions to be singletons.

Task 1A

We evaluated the Task 1A systems on both mention extraction and coreference resolution. For mention extraction, Lan *et al*[36] and Grouin *et al*[22] had an F-measure of 0.737 for the mentions that overlapped at least partially with the ground truth; Lan *et al* achieved an F-measure of 0.645 for the mentions that matched the ground truth exactly (see table 5 in the online supplements). For coreference resolution, Grouin *et al* evaluated to an unweighted average F-measure of 0.699 for mentions with at least partial overlap and an unweighted average F-measure of 0.719 for mentions with exact overlap (see table 6 and table 7 in the supplements). The baseline performance on coreference resolution on Task 1A was an unweighted average F-measure of 0.417 for both mentions wit at least partial and exact overlap (see table 6).

Task 1B

Task 1B systems' results ranged from an unweighted average F-measure of 0.827 (Glinos[37]) to an unweighted average F-measure of 0.417 (Benajiba *et al*[38]). The best performing system was rule-based (Glinos, unweighted average F-measure of 0.827), followed by a hybrid system (Rink *et al*,[28] unweighted average F-measure of 0.821), and a supervised system (Cai *et al*,[31] unweighted average F-measure of 0.806). The baseline achieved an unweighted average F-measure of 0.417 on Task 1B (see table 8 and table 9 in the supplements).

Task 1C

Sixteen systems were evaluated in Task 1C i2b2 and 20 in Task 1C i2b2/UPMC. The best scoring system in Task 1C was that of Xu *et al*,[32] with an unweighted average F-measure of 0.915 for Task 1C i2b2, and 0.913 for Task 1C i2b2/UPMC (see table 8). The supervised system of Xu *et al*[32] was followed in performance by a hybrid (Rink *et al*[28]) and two rule-based systems (Yang *et al*[26] and Hinote *et al*[23]). Of the systems that were developed in the absence of UPMC data, Dai *et al*[39] outperformed some teams who did have access to UPMC data. The baseline scored an unweighted average F-measure of 0.541 and 0.548 on Task 1C i2b2 and Task 1C i2b2/UPMC, respectively.

DISCUSSION

We analyzed system outputs on the i2b2/VA corpus and made the following observations. We expect these observations would generalize to the ODIE corpus as well.

In general, token overlap was a feature used by all systems. The person category was the easiest to handle for all systems.

Overall, the rule-based systems were able to correctly resolve the coreference on both mention pairs with exact and at least partial overlap (ie, 'a hepaticojejunostomy'—'hepaticojejunostomy,' 'a 10 pound weight gain'—'weight gain'). They correctly linked most noun phrases to their correct pronominal coreferents (ie, 'her father'—'who'). In the absence of domain knowledge, most rule-based systems were unable to link coreferent

Table 6 Task 1A coreference evaluation results using unweighted average over MUC, CEAF, and B-CUBED

	Unweighted average over MUC, CEAF, and B-CUBED					
	At least partial overlap			Exact overlap		
Team	P	R	F	P	R	F
Grouin *et al*[22]	0.642	0.814	0.699	0.662	0.848	0.719
Lan *et al*[36]	0.620	0.765	0.665	0.630	0.790	0.678
Cai *et al*[31]	0.425	0.506	0.416	0.425	0.506	0.417
Baseline	0.425	0.506	0.417	0.425	0.506	0.417

F, F-measure; P, precision; R, recall.

Table 8 Task 1B and 1C coreference evaluation results using unweighted average over MUC, CEAF, and B-CUBED

	Unweighted average over MUC, CEAF, and B-CUBED		
Team	**P**	**R**	**F**
Task 1B			
Glinos[37]	0.814	0.842	0.827
Rink *et al*[28]	0.802	0.845	0.821
Cai *et al*[31]	0.773	0.850	0.806
Grouin *et al*[22]	0.769	0.848	0.802
Hinote *et al*[23]	0.758	0.855	0.798
Lan *et al*[36]	0.753	0.805	0.777
Gooch[21]	0.582	0.701	0.620
Benajiba *et al*[38]	0.425	0.506	0.417
Baseline	0.425	0.506	0.417
Task 1C i2b2			
Xu *et al*[32]	0.906	0.925	0.915
Rink *et al*[28]	0.895	0.918	0.906
Yang *et al*[26]	0.892	0.911	0.901
Hinote *et al*[23]	0.895	0.898	0.896
Cai *et al*[31]	0.882	0.894	0.888
Anick *et al*[30]	0.857	0.915	0.883
Gooch[21]	0.895	0.858	0.875
Jindal *et al*[29]	0.901	0.830	0.861
Grouin *et al*[22]	0.850	0.862	0.856
Ware *et al*[40]	0.850	0.846	0.848
Baseline	0.517	0.597	0.541
Task 1C i2b2/UPMC			
Xu *et al*[32]	0.905	0.920	0.913
Rink *et al*[28]	0.895	0.913	0.904
Yang *et al*[26]	0.890	0.905	0.897
Hinote *et al*[23]	0.900	0.891	0.895
Cai *et al*[31]	0.881	0.885	0.883
Gooch[21]	0.898	0.859	0.878
Anick *et al*[30]	0.848	0.911	0.877
Dai *et al*[39]	0.849	0.896	0.871
Jindal *et al*[29]	0.905	0.820	0.857
Grouin *et al*[22]	0.862	0.850	0.856
Baseline	0.523	0.602	0.548

F, F-measure; P, precision; R, recall.

pairs with no token overlap (ie, ‘a ct angiogram’—‘this study,’ ‘left ankle wound’—‘a small complication’), with phrase head overlap (ie, ‘amio loading’—‘amiodarone hcl’), with abbreviations (ie, ‘an attempted ercp’—‘the endoscopic retrogram cholangiopancreatogram,’ ‘cabg’—‘surgery’), with medical terms in synonymy or hyponymy relations (ie, ‘antibiotics’—’vancomycin,’ ‘aortic insufficiency’—‘aortic sclerosis’), or with physicians and their professions (ie, ‘dr. **name [zzz]’—‘his attending physician’). The rule-based systems that incorporated domain knowledge achieved correct coreference for most mentions pairs with no token overlap (ie, ‘male’—‘the patient,’ ‘a 1,770 gram male infant’—‘the patient’), with phrase head overlap (ie, ‘hydro’—‘right hydronephrosis,’ ‘a gi bleed’—‘his bleeding’), abbreviations (ie, ‘56y/o’—‘her,’ ‘acute myelogenous leukemia’—’aml’), medical terms (ie, ‘a low blood pressure’—‘hypotension,’ ‘the subgaleal bleed’—‘the subgaleal hemorrhage,’ ‘sleepiness’—’somnolence’), and misspelled mentions (ie, ‘wound care’—‘wtd woulnd care’).

In general, the hybrid systems resolved coreference correctly when the mentions presented some degree of token overlap. These systems had an advantage over the rule-based systems in correctly linking mentions that included location pointers (ie, ‘left shoulder facture’—‘shoulder fracture,’ ‘a right lower extremity cellulitis’—‘true cellulitis’). This advantage was given by the additional processing of anatomical terms.

The supervised systems had an advantage over the hybrid and rule-based systems on mentions with no token overlap (ie, ‘a hepaticojejunostomy’—‘the procedure,’ ‘agitated’—‘increased agitation’), mentions with phrase head overlap (ie, ‘accumulation of ascitic fluid’—‘the ascites’), and mentions describing professions (ie, ‘admitting physician and thoracic surgeon’—‘dr. cranka’). Xu *et al*[32] made use of world knowledge from multiple sources (Wikipedia, WordNet, Probase, Evidence, NeedleSeek) and utilized coreference cues from intrinsic document structure. Their system correctly resolved most clinical mentions with no token overlap, including abbreviations (ie, ‘cesarean section’—‘delivery,’ ‘delta ms’—‘waxing and waning mental status’). Additionally, the supervised systems correctly linked mentions containing a larger number of tokens (ie, ‘a well developed, well nourished gentleman’—‘his’).

The 2011 challenge systems complemented each other, and collectively performed close to the ground truth. We analyzed how many system chains were identical to the ground truth chains, and identified that no chains were correctly predicted by every system and 50.48% of all ground truth chains could be correctly predicted by at least one system. Overall, 77.75% of the ground truth chains were correctly predicted by the collective efforts of all systems. In addition to the chains identical to the ground truth, the systems also predicted partially correct chains. These partially correct chains would either miss mentions or contain incorrect mentions. In order to obtain a more detailed analysis of the systems’ prediction accuracy, we performed a pairwise comparison of the system mention pairs and the ground truth mention pairs. We identified that 95.07% of the ground truth mention pairs could be correctly predicted by at least one system, and 98.92% of mention pairs could be correctly predicted by the collective efforts of all systems. Only 1.24% of the ground truth mention pairs could be correctly predicted by every system. The correct cases of coreference that all systems identified presented some degree of token overlap. The more challenging coreference cases presented no token overlap or were based on a clinical relationship (ie, ‘mild changes’—‘worsening dementia,’ ‘minimally displaced, comminuted fractures of the left c7 and t1 transverse processes’—‘rib fx’). These cases required additional external knowledge sources (ie, ‘squamous cell carcinoma’—‘stage t2 n0,’ ‘solu-medrol’—‘the iv steroids’), represented meaning distortions caused by the clinical de-identifier (ie, ‘reke, atota s’—‘she,’ ‘**name [www xxx], m.d.’—‘I’), or included misspellings (ie, ‘yeast in the urine’—‘yest’). None of the systems were able to identify coreference pairs involving metaphorical expression (ie, ‘pins and needles from the knees’—‘neuropathic type pain’).

The baseline achieved an unweighted average F-measure of 0.541 on the i2b2/VA corpus and 0.417 on the ODIE corpus. These numbers indicate the abundance of singletons in our corpora, where a system that predicts no coreference chains achieves an unweighted average F-measure which is greater than zero. However, the gains of the 2011 challenge systems over this baseline indicates that the systems were able to identify true chains and make a contribution to the coreference resolution task.

CONCLUSION

The 2011 i2b2/VA workshop on NLP challenges for clinical records focused on coreference in clinical records. Twenty teams from nine countries participated in this challenge. In general, the best performing systems incorporated domain knowledge, extracted coreference cues from the structure of the clinical records, and created dedicated modules for person concepts and for pronoun concepts. The coreference challenge results show that the current state-of-the-art medical coreference resolution systems perform well in solving coreference across all the semantic categories, but face difficulties in solving coreference for cases that require domain knowledge. More advanced incorporation of domain knowledge remains a challenge that would benefit from future research.

Acknowledgments The authors would like to thank Ying Suo and Matthew Maw for their help with data management and analysis of manual annotations data.

Contributors OU is the primary author and was instrumental in all aspects of the preparation and organization of the coreference resolution challenge from data to workshop. AB helped organize the coreference challenge, led the data analysis, and co-wrote and edited the manuscript. BS was co-lead of the coreference resolution challenge, and along with SS and TF managed the annotation and preparation of the i2b2/VA corpus for the coreference challenge. JP provided organization insights and feedback throughout challenge organization.

Funding The 2011 i2b2/VA challenge and the workshop are funded in part by grant number 2U54LM008748 on Informatics for Integrating Biology and the Bedside from the National Library of Medicine. This challenge and workshop are also supported by resources and facilities of the VA Salt Lake City Health Care System with funding support from the Consortium for Healthcare Informatics Research (CHIR), VA HSR HIR 08-374 and the National Institutes of Health, National Library of Medicine under grant number R13LM010743-01.

Competing interests None.

Ethics approval This study was conducted with the approval of i2b2 and the VA.

Provenance and peer review Not commissioned; externally peer reviewed.

REFERENCES

1. **McCarthy JF,** Lehnert WG. Using decision trees for coreference resolution. *40th International Joint Conference on Artificial Intelligence*. Montreal, Canada: IJCAII, 1995:1050—5.
2. **Haghighi A,** Klein D. Simple coreference resolution with rich syntactic and semantic features. *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. East Stroudsburg, PA, USA: Association for Computational Linguistics, 2009;**3**:1152—61.
3. **Gasperin C,** Briscoe T. Statistical anaphora resolution in biomedical texts. *Proceedings of the 22nd International Conference on Computational Linguistics*. East Stroudsburg, PA, USA: Association for Computational Linguistics, 2008;**1**:257—64.
4. **Bodnari A,** Szolovits P, Uzuner O. MCORES: A system for noun-phrase coreference resolution for clinical records. *J Am Med Inform Assoc*. Forthcoming, August 2012.
5. **Iglesias JE,** Rocks K, Jahanshad N, *et al*. Tracking medication information across medical records. *AMIA Annu Symp Proc* 2009;**2009**:266—70.
6. **Coden A,** Savova G, Sominsky I, *et al*. Automatically extracting cancer disease characteristics from pathology reports into a Disease Knowledge Representation Model. *J Biomed Inform* 2009;**42**:937—49.
7. **Zheng J,** Chapman WW, Crowley RS, *et al*. Coreference resolution: a review of general methodologies and applications in the clinical domain. *J Biomed Inform* 2011;**44**:1113—22.
8. **He T.** *Coreference Resolution on Entities and Events for Hospital Discharge Summaries*. Cambridge: MIT, 2007.
9. **Uzuner O,** Luo Y, Szolovits P. Evaluating the state-of-the-art in automatic de-identification. *J Am Med Inform Assoc* 2007;**14**:550—63.
10. **Uzuner O,** Goldstein I, Luo Y. Identifying patient smoking status from medical discharge summaries. *J Am Med Inform Assoc* 2008;**15**:14—24.
11. **Uzuner O.** Recognizing obesity and co-morbidities in sparse data. *J Am Med Inform Assoc* 2009;**16**:561—70.
12. **Uzuner O,** Solti I, Cadag E. Extracting medication information from clinical text. *J Am Med Inform Assoc* 2010;**17**:514—18.
13. **Uzuner O,** South B, Shen S, *et al*. 2010 i2b2/VA challenge on concepts, assertions, and relations in clinical texts. *J Am Med Inform Assoc* 2011;**18**:552—6.
14. **Pradhan S,** Ramshaw L, Marcus M, *et al*. CoNLL-2011 shared task: modeling unrestricted coreference in OntoNotes. *Proceedings of the Fifteenth Conference on Computational natural language learning*. East Stroudsburg, PA, USA: Association of Computational Linguistics, 2011:1—27.
15. **Savova G,** Chapman WW, Zheng J, *et al*. Anaphoric relations in the clinical narrative: corpus creation. *J Am Med Inform Assoc* 2011;**18**:459—65.
16. **Ogren P,** Savova G, Chut C. Constructing evaluation corpora for Automated clinical Named entity Recognition. *Proceedings of the 6th International language resources and evaluation*. Paris, France: European Language Resources Association (ELRA), 2008:3134—50.
17. **Vilain M,** Burger J, Aberdeen J, *et al*. A model-theoretic coreference scoring scheme. *Proceedings of the 6th Conference on Message Understanding (MUC)*. East Stroudsburg, PA, USA: Association for Computational Linguistics, 1995: 45—52.
18. **Bagga A,** Baldwin B. Algorithms for scoring coreference chains. *The First International Conference on Language Resources and Evaluation Workshop on Linguistics Coreference*. Granada, Spain, 1998:563—6.
19. **Luo X.** On coreference resolution performance metrics. *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*. East Stroudsburg, PA, USA: Association for Computational Linguistics, 2005:28—36.
20. **Noreen E.** *Computer Intensive Methods for Testing Hypotheses: an Introduction*. New-York: Wiley-Interscience, 1989.
21. **Gooch P.** Coreference resolution in clinical discharge summaries, progress notes, surgical and pathology reports: a unified lexical approach. Proceedings of the 2011 i2b2/VA/Cincinnati Workshop on Challenges in Natural Language Processing for Clinical Data. Boston, MA, USA: i2b2, 2011.
22. **Grouin C,** Dinarelli M, Rosset S, *et al*. Coreference resolution in clinical reports- the LIMSI participation in the i2b2/VA 2011 challenge. Proceedings of the 2011 i2b2/VA/Cincinnati Workshop on Challenges in Natural Language Processing for Clinical Data. Boston, MA, USA: i2b2, 2011.
23. **Hinote D,** Ramirez C, Chen P. A comparative study of co-refernece resolution in clinical text. Proceedings of the 2011 i2b2/VA/Cincinnati Workshop on Challenges in Natural Language Processing for Clinical Data. Boston, MA, USA: i2b2, 2011.
24. **Fellbaum C.** WordNet.*An Electronic Lexical Database.* Cambridge, MA: MIT Press, 1998:422.
25. **Aronson AR.** Effective mapping of biomedical text to the UMLS Metathesaurus: the metamap program. *Proc AMIA Symp* 2001:17—21.
26. **Yang H,** Willis A, Ad Roeck, *et al*. A system for coreference resolution in clinical documents. Proceedings of the 2011 i2b2/VA/Cincinnati Workshop on Challenges in Natural Language Processing for Clinical Data. Boston, MA, USA: i2b2, 2011.
27. **Jonnalagadda S,** Li D, Sohn S, *et al*. Coreference analysis in clinical notes: a multi pass sieve with alternate anaphora resolution modules. Proceedings of the 2011 i2b2/VA/Cincinnati Workshop on Challenges in Natural Language Processing for Clinical Data. Boston, MA, USA: i2b2, 2011.
28. **Rink B,** Harabagiu S. A supervised multi-pass sieve approach for resolving coreference in clinical records. Proceedings of the 2011 i2b2/VA/Cincinnati Workshop on Challenges in Natural Language Processing for Clinical Data. Boston, MA, USA: i2b2, 2011.
29. **Jindal P,** Roth D. Using domain knowledge and domain-inspired discourse model for coreference resolution for clinical narratives. Proceedings of the 2011 i2b2/VA/Cincinnati Workshop on Challenges in Natural Language Processing for Clinical Data. Boston, MA, USA: i2b2, 2011.
30. **Anick P,** Hong P, Xue N, *et al*. Coreference resolution for electronic medical records. Proceedings of the 2011 i2b2/VA/Cincinnati Workshop on Challenges in Natural Language Processing for Clinical Data. Boston, MA, USA: i2b2, 2011.
31. **Cai J,** Mujdricza E, Hou Y, *et al*. Weakly supervised graph-based coreference resolution for clinical texts. Proceedings of the 2011 i2b2/VA/Cincinnati Workshop on Challenges in Natural Language Processing for Clinical Data. Boston, MA, USA: i2b2, 2011.
32. **Xu Y,** Liu J, Wu J, *et al*. EHUATUO: a mention-pair coreference system by exploiting document intrinsic latent structures and world knowledge in discharge summaries: 2011 i2b2 challenge. Proceedings of the 2011 i2b2/VA/Cincinnati Workshop on Challenges in Natural Language Processing for Clinical Data. Boston, MA, USA: i2b2, 2011.
33. **Wikimedia Foundation I.** *Wikipedia, the Free Encyclopedia*. 2004.
34. **Song Y,** Wang H, Wang Z, *et al*. Short text conceptualization using a probabilistic knowledgebase. *Proceedings of 22nd International Joint Conference on Artificial Intelligence*. Menlo Park, CA, USA: IJCAI/AAAI, 2011:2320—36.
35. **Lee T,** Wang Z, Wang H, *et al*. Web scale taxonomy cleansing. Proceedings of the VLDB Endowment. Seattle, WA, USA: VLDB Endowment, 2011:1295—306.
36. **Lan M,** Zhao J, Zhang K, *et al*. Comparative investigation on learning-based and rule-based approaches to coreference resolution in clinic domain: a case study in i2b2 challenge 2011 Task 1. Proceedings of the 2011 i2b2/VA/Cincinnati Workshop on Challenges in Natural Language Processing for Clinical Data. Boston, MA, USA: i2b2, 2011.
37. **Glinos D.** A search based method for clinical text coreference resolution. Proceedings of the 2011 i2b2/VA/Cincinnati Workshop on Challenges in Natural Language Processing for Clinical Data. Boston, MA, USA: i2b2, 2011.
38. **Benajiba Y,** Shaw J. A SVM-based coreference resolution syste based on Philips information extraction. Proceedings of the 2011 i2b2/VA/Cincinnati Workshop on Challenges in Natural Language Processing for Clinical Data. Boston, MA, USA: i2b2, 2011.
39. **Dai H-J,** Wu C-Y, Chen C-Y, *et al*. Co-reference resolution of the medical concepts in the patient discharge summaries. Proceedings of the 2011 i2b2/VA/Cincinnati Workshop on Challenges in Natural Language Processing for Clinical Data. Boston, MA, USA: i2b2, 2011.
40. **Ware H,** Mullet C, Jagannathan V, *et al*. Coreference resolution of concepts in clinical documents. Proceedings of the 2011 i2b2/VA/Cincinnati Workshop on Challenges in Natural Language Processing for Clinical Data. Boston, MA, USA: i2b2, 2011.