

MCORES: a system for noun phrase coreference resolution for clinical records

Andreea Bodnari,¹ Peter Szolovits,¹ Özlem Uzuner²

► Additional data are published online only. To view these files please visit the journal online (<http://dx.doi.org/10.1136/amiajnl-2011-000591>).

¹MIT, CSAIL, Cambridge, Massachusetts, USA

²Department of Information Studies, University at Albany, SUNY, Albany, New York, USA

Correspondence to

Andreea Bodnari, MIT CSAIL, The Stata Center (Bldg. 32), 32 Vassar St #252, Cambridge, MA 02139, USA; andreeab@mit.edu

Received 8 September 2011

Accepted 3 February 2012

Published Online First

14 March 2012

ABSTRACT

Objective Narratives of electronic medical records contain information that can be useful for clinical practice and multi-purpose research. This information needs to be put into a structured form before it can be used by automated systems. Coreference resolution is a step in the transformation of narratives into a structured form.

Methods This study presents a medical coreference resolution system (MCORES) for noun phrases in four frequently used clinical semantic categories: persons, problems, treatments, and tests. MCORES treats coreference resolution as a binary classification task. Given a pair of concepts from a semantic category, it determines coreferent pairs and clusters them into chains. MCORES uses an enhanced set of lexical, syntactic, and semantic features. Some MCORES features measure the distance between various representations of the concepts in a pair and can be asymmetric.

Results and Conclusion MCORES was compared with an in-house baseline that uses only single-perspective 'token overlap' and 'number agreement' features. MCORES was shown to outperform the baseline; its enhanced features contribute significantly to performance. In addition to the baseline, MCORES was compared against two available third-party, open-domain systems, RECONCILE_{ACL09} and the Beautiful Anaphora Resolution Toolkit (BART). MCORES was shown to outperform both of these systems on clinical records.

Narratives in electronic medical records present information about the patient's health but are not directly available to clinical automated systems. Natural language processing (NLP) technologies can facilitate the integration of narratives into automated systems by extracting structured information (eg, names of persons, treatments) from narratives, identifying negation and uncertainty, and determining the relations between concepts (eg, LabScanner¹ and MedLEE²). A key NLP task, coreference resolution, determines whether two concepts are coreferent, ie, linked by an 'identity' or 'equivalence' relation. For example, in the sentence 'A complete blood count showed..., but the test did not...', 'the test' and 'a complete blood count' are equivalent because they refer to the same entity. Using the results of NLP analysis for clinical applications, whether these are in decision support, quality assessment, or epidemiological modeling, relies on deriving an accurate, non-redundant set of facts from the narratives. Correctly determining coreference helps to tie together different statements made about a single entity and helps to keep multiple entities that might be confused for each other distinct.

In this study, we focused on coreference resolution of noun phrases in the clinical domain. We defined mentions of concepts as markables and considered pairs of markables for coreference. We defined an ordered pair of markables I–J as a coreference candidate; in such a pair, I is the antecedent, and J is the anaphor. Groups of coreferent markables create chains, for example, I–J–K is a chain consisting of coreferent pairs I–J and J–K.

In the NLP literature, coreference resolution focused primarily on the newspaper³ and biomedical corpora,⁵ leaving the clinical corpora relatively unexplored.^{6–8} Inspired by the work of He⁹ we developed a medical coreference resolution system (MCORES) that targets noun phrases in the clinical domain, and showed that a rich set of features can be beneficial to coreference resolution. In addition, we compared MCORES with two existing third-party, open-domain (ie, not domain-specific) coreference resolution systems as a way of benchmarking it against the state-of-the-art. We found that MCORES outperforms these systems on clinical records.

RELATED WORK

MCORES was inspired by work conducted in open-domain NLP. In general, research on coreference resolution in open-domain NLP made use of annotated corpora^{10–13} and created exemplary rule-based and machine learning systems.

One of the seminal coreference resolution systems, RESOLVE, was designed by McCarthy and Lehnert.⁵ RESOLVE tackled coreference resolution of noun phrases in four steps: pair creation, feature set determination, learning, and clustering. RESOLVE focused on four semantic types: organizations, facilities, persons, and products-services. It applied the C4.5¹⁴ decision tree algorithm to classify pairs as coreferent and then clustered coreferent pairs into chains, achieving an average F-measure of 0.858 over coreference chains.

Soon *et al*¹⁵ extended RESOLVE to pronouns. They expanded RESOLVE's feature set, and modified RESOLVE's pair creation and clustering steps, achieving an F-measure of 0.626 on the Message Understanding Conference 6 (MUC-6) corpus compared with RESOLVE's F-measure of 0.472 on the same corpus. Versley *et al*¹⁶ implemented the algorithm of Soon *et al*¹⁵ in the Beautiful Anaphora Resolution Toolkit (BART). On the MUC-7 corpus, BART gave a best precision of 0.741 using the support vector machine (SVM) linear classifier, and a best recall of 0.563 using the maximum entropy classifier.

Ng and Cardie¹⁷ improved the system of Soon *et al*¹⁵ by modifying their feature set and the pair

creation algorithm, and achieved an F-measure of 0.704 on MUC-6 and 0.634 on MUC-7.

Yang¹⁸ extended RESOLVE's feature set in order to examine how the different methods of measuring the token overlap affect coreference resolution. His system achieved a best precision of 0.697 and a best recall of 0.714. Castaño *et al*¹⁹ deviated from RESOLVE's framework and focused on sortal and pronominal coreference resolution on MEDLINE abstracts. For sortal coreference resolution, they used the UMLS Metathesaurus²⁰ and MetaMap to identify biomedical markables and their semantic types. They achieved a precision of 0.733 and a recall of 0.700.

Son *et al*²¹ studied the coreference of findings of lung masses in radiology documents. Their system incorporated domain knowledge (eg, mass location, quantity, size, calcification pattern) and achieved a 0.672 MUC F-measure. Yangy *et al*²² solved coreference resolution by exploring the relationship between noun phrases and coreference clusters. Their system achieved an F-measure of 0.817 on MEDLINE abstracts.

Stoyanov *et al*²³ modeled RECONCILE_{ACLO9} after the state-of-the-art system of Ng and Cardie.¹⁷ They used a set of 76 features and applied the perceptron learning algorithm for classification. A single-link algorithm was used for clustering. RECONCILE_{ACLO9}²⁴ outperformed the systems of Soon *et al*¹⁵ and Ng and Cardie¹⁷ with a 0.712 MUC F-measure on MUC-6 and a 0.629 MUC F-measure on MUC-7.

Open-domain coreference resolution systems paved the way for coreference resolution in clinical records. Inspired by these systems, we developed MCORES for noun phrase coreference resolution in clinical records.

PROBLEM DEFINITION

We focused on noun phrases that fall under four semantic categories: persons, problems, treatments, and tests. We observed the importance of token overlap and number agreement features for successful open-domain coreference resolution. Given the observations from the open-domain:

1. We tested the hypothesis that a coreference resolution system with an enriched feature set would outperform a baseline that contained only the token overlap and number agreement features.
2. We measured the performance of MCORES against available third-party, open-domain coreference resolution systems as a way of putting its results into perspective in relation to state-of-the-art systems.
3. We evaluated the value added by various features towards coreference resolution in clinical records. We integrated into MCORES features that measure the distance between various representations of the markables in a pair. These measurements could be asymmetric, ie, produce different values depending on how they are measured. We evaluated MCORES' asymmetric features from multiple perspectives.

METHODS

Data

We used a corpus of de-identified clinical records that contained 230 discharge summaries from Partners Healthcare (PH) and 196 from Beth Israel Deaconess Medical Center (BIDMC). The Partners Healthcare records contained a total of 23 277 noun phrase markables, and the Beth Israel Deaconess Medical Center records contained a total of 16 072 noun phrase markables. These records²⁵ were provided by the i2b2 National Center for Biomedical Computing and were prepared for the coreference resolution track of the 2011 Shared Tasks for Challenges in NLP

for clinical data.²⁶ This study was approved by the relevant institutional review boards.

Annotations

We targeted the discovery of coreference chains for noun phrase markables in four semantic categories. The coreference chains in our data were built on gold standard markables with gold standard semantic categories.

Markables and their semantic categories

The markables in our corpus were annotated as a part of the 2010 i2b2/VA challenge on concepts, assertions, and relations.²⁷ These markables included tests, problems, and treatments;²⁸ however, for coreference resolution, persons and pronouns were added to the 2010 challenge markables.²⁶ Table 1 shows the number of annotated markables and chains per semantic category in our corpus.

Determining coreference chains

The corpus was doubly annotated for coreference given the gold standard markables and their semantic categories. The length of the annotated chains varied between two and 149 markables, with an average of two markables per chain for problems, treatments, and tests, and 13 markables per chain for persons (see supplementary table 2, available online only).

MCORES

MCORES resolves coreference in four steps that follow in the spirit of RESOLVE: pair creation, feature set determination, classification, and output clustering.

Noun phrase pair creation

MCORES creates positive training pairs only from neighboring markable pairs in a chain. Table 3 shows that MCORES created 94 914 pairs in our corpus; of these, 80 455 were non-coreferent and 14 459 were coreferent.

Feature set determination

Ng and Cardie¹⁷ showed that large feature sets help coreference resolution. We consequently built MCORES with a large feature set that included lexical, syntactic, and semantic information. Some of these features (eg, token distance, sentence-level markable overlap) are novel to the task of coreference resolution.

We observed that some of the frequently used features in coreference resolution are asymmetric, as they measure the distance between various representations of markables. We incorporated multiple perspectives of these features into MCORES: (1) the antecedent perspective assessed each feature from the vantage point of the antecedent; (2) the anaphor perspective was from the vantage point of the anaphor; (3) the

Table 1 Corpus description per institution: number of markables and chains in each semantic category, separated by institution

	Persons	Problems	Treatments	Tests	All categories
PH					
Markables	6104	6846	5219	5108	23 277
Chains	750	1368	972	492	3582
BIDMC					
Markables	49 22	5078	3109	2963	16 072
Chains	757	2033	1547	643	4980
Total					
Markables	11 026	11 924	8328	8071	39 349
Chains	1507	3401	2519	1135	8562

BIDMC, Beth Israel Deaconess Medical Center; PH, Partners Healthcare.

Table 3 Distribution of coreferent and non-coreferent pairs per semantic category over pairs containing exact overlap, partial overlap, and no overlap. Pair count in semantic category, percentage of pairs in semantic category, and percentage of coreferent pairs in semantic category.

	Persons			Problems			Treatments			Tests			Across all categories		
	Pair count	% of all pairs	% of all coreferent pairs	Pair count	% of all pairs	% of all coreferent pairs	Pair count	% of all pairs	% of all coreferent pairs	Pair count	% of all pairs	% of all coreferent pairs	Pair count	% of all pairs	% of all coreferent pairs
Exact overlap															
Coreferent	3347	21.792	36.599	984	1.959	33.538	786	3.608	41.831	206	2.734	41.118	5323	5.608	36.814
Non-coreferent	100	0.651		29	0.058		21	0.096		7	0.093		157	.165	
Partial overlap															
Coreferent	337	2.194	3.685	1353	2.693	46.115	764	3.507	40.660	239	3.172	47.705	2693	2.837	18.625
Non-coreferent	711	4.629		1217	2.423		557	2.557		317	4.208		2802	2.952	
No overlap															
Coreferent	5461	35.556	59.716	597	1.188	20.348	329	1.510	17.509	56	0.743	11.178	6443	6.788	44.560
Non-coreferent	5403	35.178		46056	91.679		19328	88.722		6709	89.050		77496	81.649	
Total															
Coreferent	9145	59.542		2934	5.840		1879	8.625		501	6.650		14459	15.234	
Non-coreferent	6214	40.458		47302	94.160		19906	91.375		7033	93.350		80455	84.766	

greedy perspective was the maximum of the antecedent and anaphor perspectives; and (4) the stingy perspective was their minimum. As an example, a multi-perspective token overlap feature first counted the number of tokens that were common to the antecedent and the anaphor; it created the antecedent perspective by normalizing this count by the number of tokens in the antecedent; it created the anaphor perspective by normalizing the count by the number of tokens in the anaphor; greedy and stingy perspectives were created by applying max and min to the antecedent and anaphor perspectives. We refer to features that do not have multiple perspectives, as well as each of the individual perspectives of multi-perspective features, as single-perspective features.

Phrase-level lexical features

MCORES included the following phrase-level lexical features:

1. Token overlap (multi-perspective): Percentage of tokens that two markables share.
2. Normalized token overlap (multi-perspective): Degree of token overlap between the normalized forms of markables. Normalization was performed by the 'norm' function of the unified medical language system (UMLS) lexical variant generator.²⁰
3. Edit distance (single-perspective): Character-level Levenshtein distance between two markables. The Levenshtein distance is symmetric and does not benefit from multiple perspectives.
4. Normalized edit distance (single-perspective): Edit distance on normalized markables.

Sentence-level lexical features

We hypothesized that two coreferent markables will probably be surrounded by similar tokens and markables, and we supplemented MCORES with sentence-level lexical information that captured context:

1. Sentence-level token overlap (multi-perspective): Percentage of tokens that were common to the antecedent's and the anaphor's sentences.
2. Filtered sentence-level token overlap (multi-perspective): Sentence-level token overlap after the sentences were filtered of stop words.
3. Left and right markable overlap (stingy and greedy perspectives only): Percentage of tokens that were common to the closest left and right markables of the antecedent and the anaphor in their respective sentences.

Syntactic features

Syntactic features of markables can help resolve coreference. MCORES included:

1. Number agreement (single-perspective): We tracked if two markables agreed in number.
2. Noun overlap (multi-perspective): We counted the number of nouns a pair of markables shared (eg, 'chest pain' could be assigned to the chain 'chest'—'upper front of body'—'thorax' or to the chain 'chronic pain in upper body'—'chest affliction' depending on the level of overlap).
3. Surname match (single-perspective): We checked if two person markables included the same surname.

Semantic features

MCORES mapped markables to UMLS concept unique identifiers (CUI) using MetaMap; it filtered these CUIs according to their UMLS scores²⁰ and the number of their UMLS semantic types, creating a set of UMLS CUIs for the antecedent (set I) and a set for the anaphor (set J). This process also created a set of UMLS semantic types for the antecedent (set S) and the anaphor (set T):

1. UMLS CUI overlap (multi-perspective): Percentage of CUIs that appeared in both I and J.

2. UMLS CUI token overlap (multi-perspective): For each pair (u, v), $u \in I, v \in J$, we counted the number of tokens common to them. We summed this count across all unique pairs (u, v) and converted it to a percentage of either I or J.
3. UMLS semantic type overlap (multi-perspective): For each pair (s, t), $s \in S, t \in T$, we counted the number of tokens common to them. We summed this count across all unique pairs (s, t) and converted it to a percentage of either S or T.
4. Anaphor UMLS semantic type (single-perspective): UMLS semantic type of the anaphor.

Miscellaneous features

MCORES supplemented its feature set with a series of single perspective attributes:

1. Token distance (single-perspective): Number of tokens between the antecedent and anaphor.
2. Markable distance (single-perspective): Number of markables of the same semantic category located between the antecedent and the anaphor.
3. All-markable distance (single-perspective): Number of markables of any semantic category located between the antecedent and the anaphor.
4. Sentence distance (single-perspective): Number of sentences separating the two markables.
5. Section match (single-perspective): Discharge summaries are organized into sections such as ‘Past medical history’, ‘History of present illness’, and ‘Hospital course’. Markables that are in similar sections can more likely corefer.
6. Section distance (single-perspective): Number of sections separating the antecedent and the anaphor.

Pair classification

MCORES used the C4.5 decision tree algorithm. We selected this algorithm for its flexibility, prediction model readability, and established track record.¹⁴

Chain creation

We used RESOLVE’s aggressive merge to cluster coreferent pairs into chains. Aggressive merge clustered a coreferent pair A–B with all coreferents linked to A or B.

Evaluation metrics

We evaluated coreference systems on pairs (see supplementary data, available online only) and on chains.

Evaluation of chains

We evaluated chains using four widely used metrics that possess different strengths: MUC,²⁹ B-Cubed (B^3),³⁰ CEAF,³¹ and BLANC.³² The MUC metric ignores recall for singleton chains (ie, chains that consist of a single markable), and favors systems that generate longer chains (ie, a system that generates a single chain of all the markables will receive 100% recall, and a fairly high precision). The B^3 metric takes singletons into account. Because multiple markables can belong to a single chain, the B^3 metric can count the same chain many times. To avoid this, the

CEAF metric aligns entities in the system response and the gold standard before evaluating performance. BLANC adjusts the Rand index³³ for coreference resolution. It takes singletons into consideration and evaluates correctly identified chains according to the number of markables they contain.

We used the unweighted average of the four metrics as a measure of coreference performance on chains and evaluated pair classification using F-measure. For details on each of the metrics, see supplementary data (available online only).

Significance testing

We used the approximate randomization test³⁴ to assess whether two system outputs were significantly different from each other. We set α to 0.05. Because we compared multiple hypotheses, we applied the Bonferroni³⁵ correction to counteract the problem of multiple comparisons. The Bonferroni adjusted α was set to 0.00045 for 111 comparisons (see table 5). See supplementary data (available online only) for details.

Systems

Comparison systems

We evaluated MCORES against a baseline that was identical to MCORES except in its feature set. The baseline employed only token overlap and number agreement as features; comparison with this baseline revealed the gain of MCORES’ features over those of the baseline.

We also compared MCORES against RECONCILE_{ACL09} and BART, two state-of-the-art, open-domain systems. Both RECONCILE_{ACL09} and BART automatically generate features from any given corpus and can therefore be applied to any domain. RECONCILE_{ACL09} and BART differ from MCORES in their classification and clustering algorithms (see table 4). Given these differences from MCORES, we present the results of these systems only as a benchmark, and mean for them to demonstrate the performance on clinical records of state-of-the-art, open-domain systems.

Last, but not least, to evaluate each of the features of MCORES, we ran it with subsets of its features. Each of the instances of MCORES run with subsets of features was referred to as sub-MCORES.

System runs

RECONCILE_{ACL09} and BART were designed to handle all markables in the corpus together, regardless of their semantic categories. We therefore evaluated MCORES, the baseline, sub-MCORES, RECONCILE_{ACL09}, and BART under this scenario:

- a. We ignored the semantic category distinction between the markables and ran each system once per corpus. This allowed each of the systems to create pairs, select features, classify pairs, and cluster pairs using all of the markables from all semantic categories in the corpus. We call these the per-corpus runs.

Table 4 Summary of the characteristics of MCORES, the baseline, RECONCILE_{ACL09}, and BART

Characteristic	MCORES	BASELINE	RECONCILE _{ACL09}	BART
Pair creation	Soon <i>et al</i> ¹⁵	Soon <i>et al</i> ¹⁵	Ng and Cardie ¹⁷	Soon <i>et al</i> ¹⁵
Classification algorithm	C4.5	C4.5	Perceptron	SVM linear
Clustering algorithm	Aggressive merge	Aggressive merge	Single-link	Closest-link
Feature count	20	2	76	11
Domain of development	Clinical domain	Generic	Open-domain NLP	Open-domain NLP
Includes token match feature	Yes	Yes	Yes	Yes
Clinical knowledge	Yes	No	No	No

Table 5 MCORES versus sub-MCORES unweighted average F-measure across MUC, B³, CEAF, and BLANC metrics

System	Per-entity runs					Per-corpus runs				
	Persons	Problems	Treatments	Tests	Across all markables	Persons	Problems	Treatments	Tests	Across all markables
MCORES	0.842	0.852	0.886	0.845	0.898	0.613	0.763	0.778	0.744	0.749
Baseline	0.663*	0.846	0.870	0.843*	0.804*	0.575*	0.775*	0.804*	0.765*	0.730
RECONCILE _{ACL09}	0.389*	0.556*	0.560*	0.604*	0.582*	0.389*	0.556*	0.560*	0.604	0.541
BART										0.612*
Feature-based sub-MCORES										
Phrase-level lexical	0.822*	0.852	0.886	0.840	0.893	0.584*	0.772	0.804*	0.633	0.729
Sentence-level lexical	0.814	0.599*	0.590*	0.685	0.834*	0.391*	0.567*	0.578	0.632	0.548
Syntactic	0.847	0.562*	0.560*	0.604*	0.842*	0.389*	0.555*	0.560*	0.604	0.530
Semantic	0.839*	0.555*	0.560*	0.604*	0.849*	0.389*	0.555*	0.560*	0.604	0.541
Miscellaneous	0.841	0.556*	0.560*	0.635*	0.824*	0.408*	0.561*	0.561*	0.602	0.554*
Perspective-based sub-MCORES										
Antecedent	0.845*	0.844	0.868	0.838	0.893	0.617	0.751*	0.759	0.737*	0.753
Anaphor	0.847*	0.845	0.875	0.836	0.893	0.629*	0.757*	0.778	0.747	0.743*
Greedy	0.849	0.854*	0.884	0.848*	0.899	0.612*	0.747	0.780	0.741*	0.749
Stingy	0.847	0.848	0.872	0.843	0.896	0.613	0.758*	0.772*	0.739*	0.747*

*Significantly different from MCORES' results, at Bonferroni-corrected alpha=0.00045.

However, given the availability of ground truth semantic category information for our markables, we also ran MCORES, the baseline, sub-MCORES, and RECONCILE_{ACL09} on individual semantic categories as follows.

b. We took the semantic category information into account and ran each system four times, once for each of the semantic categories. This allowed each of the systems to create pairs, select features, classify pairs, and cluster pairs on one semantic category at a time. We call these the per-entity runs. BART's design did not allow for it to be adapted for the per-entity runs.

We evaluated the runs:

- ▶ across all markables at the same time, without a distinction in semantic category, and
- ▶ on individual semantic categories, focusing on markables of one semantic category at a time.

Given its design, we evaluated BART across all markables only. All systems were cross-validated (10-fold) on our data.

RESULTS

Evaluating the task

We evaluated the difficulty of coreference resolution in our corpus by checking the degree of token overlap between the markables in pairs.

Table 3 shows that approximately 30–40% of coreferent pairs in each of the semantic categories in our corpus present exact overlap (markables overlap in their entirety). At least 40% of problem, treatment, and test coreferent pairs present partial overlap (markables share at least one token and, as supplementary table 2 (available online only) shows, the average markable length is 5.28 across all semantic categories). In addition, on average 1% of non-coreferent pairs show exact overlap (eg, two x-rays taken on two different days) and between 2% and 5% of non-coreferent pairs show at least partial overlap.

Evaluating the systems

Given these data, we then investigated the three questions from the problem definition section. We repeat these questions here for convenience:

1. We tested the hypothesis that MCORES would outperform a baseline that used only the token overlap and number agreement features (see section MCORES vs the baseline).

2. We measured the performance of MCORES against available third-party, open-domain coreference resolution systems as a way of putting its results into perspective in relation to state-of-the-art systems (see section MCORES vs the third-party systems).

3. We evaluated MCORES features, including asymmetric ones, for coreference resolution in clinical records, (see section Evaluation of features and perspectives).

MCORES versus the baseline

MCORES performance

MCORES performed well on both pairs and chains (see table 5 for chains and supplementary table 6, available online only, for pairs). It achieved an unweighted average F-measure of 0.749 for per-corpus and 0.898 for per-entity runs when evaluated on chains across all markables. MCORES' best per-entity run, with an unweighted average F-measure of 0.886, was on treatments (see table 5). MCORES' performance on individual semantic categories was lower on the per-corpus runs than on the per-entity runs.

Baseline comparison on chains

MCORES significantly outperformed the baseline across all markables in per-entity runs; its per-corpus unweighted average F-measure was 0.749 versus 0.730 of the baseline; its per-entity unweighted average F-measure was 0.898 versus 0.804 of the baseline (see table 5). Extended results are included in supplementary table 7 (available online only).

Baseline comparison on pairs

MCORES also outperformed the baseline in pair classification (see supplementary table 6, available online only). The per-entity run of MCORES performed particularly well on pair classification. It had an F-measure of 0.824 across all markables with best performance on persons and lowest performance on problems (for details see supplementary text, available online only).

We evaluated pair classification on exact token overlap, partial token overlap, and no token overlap pairs separately. MCORES outperformed the baseline on all pair types for per-corpus and for per-entity runs when evaluated across all markables (see supplementary table 8, available online only). When evaluated on individual semantic categories, MCORES outperformed the baseline on some and was outperformed on others for both the per-entity and per-corpus runs.

MCORES versus the third-party systems

We evaluated MCORES against RECONCILE_{ACL09} and BART. Table 5 shows that MCORES outperformed RECONCILE_{ACL09} across all markables on per-entity runs. It outperformed RECONCILE_{ACL09} on persons, problems, and treatments on both per-entity and per-corpus runs. MCORES outperformed BART across all markables on the per-corpus runs.

Evaluation of features and perspectives

To evaluate MCORES' lexical, syntactic, semantic, and miscellaneous features, we ran it with each group of features separately. Each of the runs with subsets of MCORES' features is referred to as feature-based sub-MCORES.

Table 5 shows that MCORES outperformed the feature-based sub-MCORES (except for the phrase-level lexical sub-MCORES) on per-entity runs when evaluated across all markables (see table 5). It outperformed the miscellaneous sub-MCORES on per-corpus runs when evaluated across all markables. Analysis on individual semantic categories of per-entity runs shows that MCORES performed better or as well as all feature-based sub-MCORES on all semantic categories. Analysis on individual semantic categories of per-corpus runs shows that MCORES outperformed all feature-based sub-MCORES on persons, it outperformed all but one feature-based sub-MCORES (phrase-level lexical sub-MCORES) on problems, it outperformed all but one feature-based sub-MCORES (sentence-level lexical sub-MCORES) on treatments, and was comparable to all feature-based sub-MCORES on tests.

To measure the value of multi-perspective features, we evaluated each of the individual perspectives against MCORES. Each of the runs with individual perspectives is referred to as perspective-based sub-MCORES. The results of comparing antecedent, anaphor, greedy, and stingy perspective-based sub-MCORES with MCORES are shown in table 5. Comparing the perspective-based sub-MCORES with each other, we find that, in the per-entity runs, greedy perspective sub-MCORES gives the best performance; in the per-corpus runs, antecedent perspective sub-MCORES gives the best performance.

DISCUSSION

In general, MCORES outperformed the baseline and the third-party systems across all markables. In the per-entity runs, MCORES significantly outperformed the baseline on persons and tests. In the per-corpus runs, MCORES significantly outperformed the baseline on persons, problems, treatments, and tests.

Analysis of system outputs revealed several patterns. The gain of MCORES in persons over the baseline came from its ability to link markables with no token overlap (ie, 'patient'—'Kulrine, ryyege n'); on tests, the baseline showed a strong disadvantage by linking unrelated markables with partial token overlap (ie, the baseline incorrectly links 'the mri on admission'—'the hct on admission'). In general the baseline overgenerated chains; this was coincidentally to its advantage on the less prevalent classes, such as treatments. However, we expect that this advantage would disappear as the data set grows.

Both MCORES and the phrase-level lexical sub-MCORES outperformed the third-party systems (see table 5). The third-party systems generally underpredicted the true coreference pairs; this was accounted for by their pair creation methods. Despite the performance displayed by the phrase-level lexical sub-MCORES with no clinical knowledge, clinical knowledge played a role in the gain of MCORES over RECONCILE_{ACL09}. For example, pairs such as 'right basilar atelectasis'—'atelectasis'

were correctly classified by MCORES but were missed by RECONCILE_{ACL09}, BART, and the phrase-level lexical sub-MCORES.

Phrase-level lexical sub-MCORES performed similarly to MCORES on most individual semantic categories and across all markables. However, much like the baseline, phrase-level lexical sub-MCORES tended to link markable pairs incorrectly with partial overlap and mostly generated shorter chains than the gold standard. We expect that as the data grow, MCORES will also gain over this sub-MCORES on the individual semantic categories.

We found that individual perspective-based sub-MCORES performed similarly to each other and to MCORES when evaluated across all markables (see table 5). While some of their differences were statistically significant, the observed differences may not justify the additional model complexity and a greedy perspective sub-MCORES may in general be sufficient for our application.

MCORES did not always outperform its competitors. For example, it did not significantly outperform the perspective-based sub-MCORES and the phrase-level sub-MCORES on the per-entity runs. It also did not show significantly better performance on treatments, tests, and across all markables on the per-corpus runs. Yet, there is no single system that could outperform MCORES on both the per-corpus and per-entity runs across all markables and across all semantic categories.

MCORES generated typical errors for each semantic category; for example, it failed to classify misspelled person pairs. False positives for problems were generated by the inability to distinguish between newly arisen and recurring events (eg, pneumonia occurring at different dates vs an incurable disease such as AIDS). Treatment false positives were encountered when medications with the same name (but different routes of administration) did not corefer. Test errors occurred because many test pairs that exhibited exact token overlap did not corefer. For all semantic categories, false negatives mainly came from markables with no token overlap; a shortcoming that could be remedied by additional world knowledge, or by better filtering the knowledge provided by the UMLS (eg, the system should infer that 'infection' and 'communicable disease' referred to the same entity).

Last, but not least, this paper relied on markables and semantic categories to be annotated before resolving coreference so that we could focus on coreference resolution without having to worry about the noise that could be introduced by automatic processes. Obtaining such corpora is difficult; however, manual markable and semantic category information can be replaced by their automated counterparts at the expense of some system performance.

CONCLUSION

We presented MCORES, a coreference resolution system that is modeled after RESOLVE but includes significant expansions to the original feature set. Our evaluation of coreference resolution in the clinical domain found token overlap to be a very helpful, but insufficient, feature that can overgenerate chains. With a feature set that enhances token overlap with lexical, syntactic, and semantic information, we showed that MCORES outperformed an in-house baseline and two third-party systems, improving coreference resolution on clinical records.

Contributors AB is the primary author and was instrumental in designing and developing the work and performed data analyses. PS and OU are the principal investigators for the grant involving the secondary use of clinical data. OU co-designed the experiments, led the data analysis, provided expertise in machine learning, and

co-wrote and edited the manuscript. PS provided expertise in data analysis and reviewed and edited the manuscript. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NLM, NHLBI, NIH, or ONC.

Funding This project was supported in part by award number 2U54LM008748 from the National Institutes of Health (NIH)/National Library of Medicine (NLM) and co-founded by the National Heart, Lung and Blood Institute (NHLBI), and by contract number 90TRO002 (SHARP—Secondary Use of Clinical Data) from the Office of the National Coordinator (ONC) for Health Information Technology and by contract number EB001659 from the National Institute of Biomedical Imaging and Bioengineering.

Competing interests None.

Ethics approval This study was conducted with the approval of the institutional review boards of Partners Health Care, MIT, SUNY at Albany.

Provenance and peer review Not commissioned; externally peer reviewed.

Data Sharing Statement Data are available from i2b2.org/NLP.

REFERENCES

1. Post A, Harrison J. An enhanced framework for pattern detection in clinical laboratory data. *Proceedings of the AMIA Symposium*. Vol 2002. Boston, MA, USA: American Medical Informatics Association, 2002:1134.
2. Friedman C, Alderson PO, Austin JH, et al. A general natural-language text processor for clinical radiology. *J Am Med Inform Assoc* 1994;1:161–74.
3. McCarthy JF, Lehnert WG. Using decision trees for coreference resolution. *Proceedings of the 14th International Joint Conference on Artificial Intelligence*. Vol 2. Montreal, Canada: Morgan-Kaufmann, 1995:1050–5.
4. Haghighi A, Klein D. Simple coreference resolution with rich syntactic and semantic features. *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Vol 3. Singapore: Association for Computational Linguistics, 2009:1152–61.
5. Gasperin C, Briscoe T. Statistical anaphora resolution in biomedical texts. *Proceedings of the 22nd International Conference on Computational Linguistics*. Vol 1. Manchester, UK: Association for Computational Linguistics, 2008:257–64.
6. Iglesias JE, Rocks K, Jahanshad N, et al. Tracking medication information across medical records. *Proceedings of the AMIA Annual Symposium*. Vol 2009. San Francisco, CA, USA: American Medical Informatics Association, 2009:266–70.
7. 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.
8. Zheng J, Chapman WW, Crowley RS, et al. Coreference resolution: a review of general methodologies and applications in the clinical domain. *J Biomed Inform*. Published Online First: 12 August 2011. doi:10.1016/j.jbi.2011.08.006
9. He T. *Coreference Resolution on Entities and Events for Hospital Discharge Summaries*. Cambridge: MIT, 2007.
10. Chinchor NA. Overview of MUC-7/MET-2. *Proceedings of the 7th Message Understanding Conference (MUC-7)*. Fairfax, VA, USA, 1998. http://www-nlpir.nist.gov/related_projects/muc/proceedings/muc_7_proceedings/overview.html
11. Grishman R, Sundheim B. Message Understanding Conference-6: a brief history. *Proceedings of the 16th Conference on Computational Linguistics*. Vol 1. Copenhagen, Denmark: Association for Computational Linguistics, 1996:466–71.
12. Doddington G, Mitchell A, Przybocki M, et al. The Automatic Content Extraction (ACE) program—tasks, data, and evaluation. *Proceedings of the 4th International Conference on Language Resources and Evaluation*. Vol 4. Lisbon, Portugal: European Language Resource Association, 2004:837–40.
13. Recasens M, Martí T, Taulé M, et al. SemEval-2010 Task 1: coreference resolution in multiple languages. *Proceedings of the Semantic Evaluations Workshop: Recent Achievements and Future Directions*. Vol 1. Boulder, CO, USA: Association for Computational Linguistics, 2009:70–5.
14. Quinlan JR. *C4.5: Programs for Machine Learning*. San Mateo, CA: Morgan Kaufmann, 1993.
15. Soon WM, Ng HT, Lim DC. A machine learning approach to coreference resolution of noun phrases. *Comput Ling* 2001;27:521–44.
16. Versley Y, Ponzetto S, Poesio M, et al. BART: A modular toolkit for coreference resolution. *Proceedings of the 6th International Language Resources and Evaluation*. Vol 1. Marrakech, Morocco: European Language Resources Association, 2008:9–12.
17. Ng V, Cardie C. Improving machine learning approaches to coreference resolution. *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*. Vol 1. Philadelphia, PA, USA: Association for Computational Linguistics, 2002:104–11.
18. Yang H. Automatic extraction of medication information from medical discharge summaries. *J Am Med Inform Assoc* 2010;17:545–8.
19. Castaño J, Zhang J, Pustejovsky J. Anaphora resolution in biomedical literature. In *Proceedings of the International Symposium on Reference Resolution*. 2002.
20. Aronson AR. Effective mapping of biomedical text to the UMLS Metathesaurus: the metamap program. *Proceedings of the AMIA Annual Symposium*. Vol 2001. Washington, DC, USA: American Medical Informatics Association, 2001:17–21.
21. Son R, Taira R, Kangaroo H. Inter-document coreference resolution of abnormal findings in radiology documents. *Stud Health Technol Inform* 2004;107:1388–92.
22. Yangy X, Su J, Zhou G, et al. An NP-cluster based approach to coreference resolution. *Proceedings of the 20th International Conference on Computational Linguistics*. Vol 1. Geneva, Switzerland: Association for Computational Linguistics, 2004:226–32.
23. Stoyanov V, Cardie C, Gilbert N, et al. Coreference resolution with reconcile. *Proceedings of the ACL 2010 Conference Short Papers*. Vol 1. Uppsala, Sweden: Association for Computational Linguistics, 2010:156–61.
24. Stoyanov V, Gilbert N, Cardie C, et al. Conundrums in noun phrase coreference resolution: making sense of the state-of-the-art. *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*. Vol 2. Suntec, Singapore: Association for Computational Linguistics, 2009:656–64.
25. Saeed M, Villarroel M, Reisner AT, et al. Multiparameter Intelligent Monitoring in Intensive Care II (MIMIC-II): a public-access intensive care unit database 2011. *Crit Care Med* 2011;39:952–60.
26. i2b2. 2011 i2b2/VA/Cincinnati Shared-Task and Workshop. 2011. <https://www.i2b2.org/NLP/Coreference/> (accessed Aug 2011).
27. 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.
28. Uzuner O, Solti I, Cadag E. Extracting medication information from clinical text. *J Am Med Inform Assoc* 2010;17:514–18.
29. Vilain M, Burger J, Aberdeen J, et al. A model-theoretic coreference scoring scheme. *Proceedings of the 6th Conference on Message Understanding (MUC-6)*. Vol 1. Columbia, MD, USA: Morgan Kaufmann, 1995:42–52.
30. Bagga A, Baldwin B. Algorithms for scoring coreference chains. *The First International Conference on Language Resources and Evaluation Workshop on Linguistics Coreference*. Vol 1. Granada, Spain: Citeseer, 1998:563–6.
31. Luo X. On coreference resolution performance metrics. *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*. Vol 1. Vancouver, British Columbia, Canada: Association for Computational Linguistics, 2005:28–36.
32. Recasens M, Hovy E. BLANC: Implementing the Rand index for coreference evaluation. *Nat Lang Eng* 2010;17:485–510.
33. Rand WM. Objective criteria for the evaluation of clustering methods. *J Am Stat Assoc* 1971;66:846–50.
34. Noreen E. *Computer Intensive Methods for Testing Hypotheses: An Introduction*. New York: Wiley-Interscience, 1989.
35. Miller R. *Simultaneous Statistical Inference*, 2nd edn. New York: Springer Verlag, 1982.