Downloaded from jamia.bmj.com on August 4, 2012 - Published by group.bmj.com

A supervised framework for resolving coreference in clinical records

Bryan Rink, Kirk Roberts, Sanda M Harabagiu

Human Language Technology Research Institute, University of Texas at Dallas, Richardson, Texas, USA

Correspondence to
Bryan Rink, University of Texas at Dallas, PO Box 830688; MS EC31, Richardson, TX 75083-0688, USA;
bryan@hlt.utdallas.edu

Received 3 January 2012
Accepted 15 April 2012

ABSTRACT

Objective A method for the automatic resolution of coreference between medical concepts in clinical records.

Materials and methods A multiple pass sieve approach utilizing support vector machines (SVMs) at each pass was used to resolve coreference. Information such as lexical similarity, recency of a concept mention, synonymy based on Wikipedia redirects, and local lexical context were used to inform the method. Results were evaluated using an unweighted average of MUC, CEAF, and B^3 coreference evaluation metrics. The datasets used in these research experiments were made available through the 2011 i2b2/VA Shared Task on Coreference.

Results The method achieved an average F score of 0.821 on the ODIE dataset, with a precision of 0.802 and a recall of 0.845. These results compare favorably to the best-performing system with a reported F score of 0.827 on the dataset and the median system F score of 0.800 among the eight teams that participated in the 2011 i2b2/VA Shared Task on Coreference. On the i2b2 dataset, the method achieved an average F score of 0.906, with a precision of 0.895 and a recall of 0.918 compared to the best F score of 0.915 and the median of 0.859 among the 16 participating teams.

Discussion Post hoc analysis revealed significant performance degradation on pathology reports. The pathology reports were characterized by complex synonymy and very few patient mentions.

Conclusion The use of several simple lexical matching methods had the most impact on achieving competitive performance on the task of coreference resolution. Moreover, the ability to detect patients in electronic medical records helped to improve coreference resolution more than other linguistic analysis.

BACKGROUND AND SIGNIFICANCE

The adoption of electronic medical records (EMRs) has enabled the use of automatic methods for analyzing, reviewing, and querying patient records. Natural language processing (NLP) technology contributes to many of the automatic analysis methods that provide a wide variety of applications. This is due to the fact that many EMRs contain an unstructured, narrative portion where important medical information is recorded. However, most NLP techniques have been traditionally developed for processing very different discourse domains (eg, news), and in very different genres (eg, financial or political). For the medical domain, NLP techniques need to take into account significant semantic information available in various medical ontologies. Moreover, the narratives are written in quite different styles than for other domains, imposing a variety of processing techniques that capture the pragmatics of clinical discourse.

As clinical narratives repeatedly refer to the same concept, special NLP techniques need to be developed to resolve references. Mentions of people (eg, the patient, the doctor), tests (eg, x-ray, blood count), treatments (eg, drugs, surgeries, therapy), and problems (eg, diabetes, atrial fibrillation) need to be resolved to specific entities. Existing NLP systems can accurately identify medical concepts in EMRs,[1] however those systems do not identify whether several concept mentions actually refer to the same concept. For example, in the sentence '*The patient's cardiovascular status has been stable throughout his CMED CSRU stay*,' the concepts '*The patient*' and '*his*' refer to the same person. In other words, these concepts are coreferential. Automatically resolving such coreferences enables the consolidation of medical information that would otherwise appear unrelated.

Coreference resolution is known to be a complex and difficult problem[2–6] because it relies on syntactic, semantic, and mostly pragmatic knowledge that is difficult to discern from narratives. However, when documents pertain to a specific domain, for example, medicine, pragmatic knowledge is replaced largely by domain-specific knowledge, which can be modeled by domain-specific concepts.

To exemplify the complexity of knowledge required for resolving coreference in clinical texts, consider the relationship between the underlined phrases involving laparoscopy (a camera-aided procedure through the abdomen) from the following example:

1. In November 2008 her doctors in Louisiana did an exploratory laparoscopy.
2. The laparoscopy saw only some minimal endometriosis.
3. The pain returned and she had a repeat laparoscopy which showed nothing.
4. She has a new-onset of pain since her surgery.

The two mentions of laparoscopy from sentences 1 and 2 refer to the same procedure. By examining the lexical overlap between the mentions, it can be inferred that they refer to the same treatment. While such an assumption often holds true, the mentions of laparoscopy from sentences 2 and 3 are actually referring to different procedures, because the mention from sentence 3 is qualified with the word 'repeat.' Thus, we need syntactic knowledge to understand that 'repeat' is an adjective modifying the concept of 'laparoscopy' and semantic knowledge to know that a 'repeat laparoscopy' refers to a second procedure.

Copyright 2012 by American Medical Informatics Association.

Downloaded from jamia.bmj.com on August 4, 2012 - Published by group.bmj.com

A different type of semantic knowledge must be used to determine that the concepts 'laparoscopy' and 'surgery' in sentences 3 and 4 are referring to the same procedure. In this instance the reader must understand that laparoscopy is a type of surgery and the doctor has referred back to the same concept using a generalization. Hence, achieving the most accurate resolution of coreference in clinical records automatically requires the incorporation of multiple forms of linguistic knowledge.

Performing automatic coreference resolution provides valuable information when extracting knowledge from clinical records. In the example sentences above, simply knowing how many laparoscopies the patient has undergone requires knowledge about which mentions of laparoscopy are referring to the same procedure and which ones are distinct. Furthermore, with coreference information, details extracted about individual concept mentions can be merged to form a more complete picture, such as the fact that a November 2008 laparoscopy revealed only minimal endometriosis.

The 2011 i2b2/VA Shared Task on Coreference focused on evaluating techniques for clinical coreference resolution, providing both a training and testing dataset. We use these datasets for evaluation and describe them further in the section on Materials and methods.

Related work

Coreference resolution has been studied for years in the NLP literature.[2 5 7 8] Approaches have included both supervised[4 9–11] and unsupervised methods.[12–15] Bengston and Roth[9] showed that a detailed focus on high quality features outperforms more sophisticated models in a supervised setting. Haghighi and Klein[12] were the first to report that a generative model which jointly models entities across multiple documents can perform at a state-of-the-art level with very little supervision. Nicolae and Nicolae[16] introduced BestCut, which treats coreference resolution as a graph cutting problem, achieving state-of-the-art performance.

A common approach to coreference resolution involves determining the best antecedent for every mention. Our approach instead makes decisions for individual pairs of concepts, possibly determining that a single concept is coreferential with many other concepts. Ng and Cardie[4] review several methods which chose a single best previous mention. These include Closest-Link, which makes pair-wise decisions about coreference, but only keeps the most recent antecedent in the final determination. Another such method is the Best-Link strategy, which generates scores for every antecedent of a mention. The mention is linked only with the highest scoring antecedent if the score is above a threshold. More recently, Raghunathan *et al*[15] report on a multiple pass sieve approach which we describe later.

Zheng *et al* provide a good review of the literature on coreference resolution in the clinical domain.[17] Wang *et al* evaluated pronominal coreference for the words 'it,' 'this,' and 'that' within 1000 sentences taken from clinical text.[18] Using a rule-based approach they achieved results ranging from 90% to 94%. He *et al*[19] studied coreference in hospital discharge summaries involving five types of entities using a supervised C4.5 decision tree classifier and a carefully selected set of features. Previous results are also available on the ODIE dataset which comprises one of our evaluation corpora. Zheng *et al* report on the results of a support vector machine (SVM)-based approach trained on syntactic, semantic, and surface features.[20] Research on biomedical literature, particularly biomedical scholarly articles, has also discussed the coreference problem. However, such articles usually focus on coreference resolution for drugs,[21] genes,[22] proteins, and bio-processes,[23] which occur less frequently in the clinical domain.

Based on previous work on coreference resolution, which suggests that supervised approaches do well when there is a sizeable training corpus available for training models, we chose to incorporate the knowledge derived from the annotations on the training data through the use of supervised classifiers in our approach.

MATERIALS AND METHODS

i2b2/VA 2011 Shared Task on Coreference

The 2011 i2b2/VA Shared Task on Coreference follows a series of annual shared tasks having different NLP focuses. The 2010 task made available a corpus of annotated concepts and their concept types. The 2011 task extends this by additionally annotating coreference between the concepts. Concepts have been annotated in two different ways: (1) according to the i2b2 guidelines, and (2) according to the ODIE[24] guidelines. The i2b2 guidelines specify five concept types: PROBLEM, TREATMENT, TEST, PERSON, and PRONOUN. The ODIE guidelines were developed independently and contain concept types such as PEOPLE, DISEASE-OR-SYNDROME, SIGN-OR-SYMPTOM, ANATOMICAL-SITE, PROCEDURE, ORGAN-OR-TISSUE-FUNCTION, LABORATORY-TEST-OR-RESULT, OTHER, and NONE.

Multiple pass sieve strategy for resolving coreference

We use a multiple pass sieve approach similar to Raghunathan *et al*.[15] This method involves multiple independent models for resolving coreference which are executed in succession. Each model (or pass) makes coreference decisions on pairs of concepts from the text. Given a pair of concepts, a model can decide either that the two concepts from a pair are coreferential or that they are not. Rather than considering all possible pairs of concepts from the text, each model has its own *selection criteria* for choosing a subset of those pairs on which to make decisions. For instance, one pass identifies coreferential pairs of mentions which are synonymous (eg, 'GERD' and 'Gastro-esophageal reflux disease'), while another pass identifies coreferential pairs of mentions whose strings are identical (eg, 'attending physician' and 'attending physician'). Unlike the approach in Raghunathan *et al*,[15] each of our passes uses a machine-learned classifier to identify which of the pairs of mentions are actually coreferential. For instance, rather than assuming that all mentions sharing the same exact text are coreferential, we train a binary classifier to make the final determination.

As seen in figure 1, coreference resolution is performed by executing the passes sequentially. Each pass makes use of the coreference decisions output by the previous passes. The final set of coreference chains is then the combination of running individual coreference passes that are specialized at identifying specific types of coreference. All of the passes use an SVM classifier provided by the LIBLINEAR library[25] with default settings. With the exception of pass 1, all of the passes share a common set of features describing properties of a pair of concepts, with some passes adding additional features. These features are used by the classifier to make a determination about whether the pair of concepts is coreferential. Each pass addresses a different coreference problem. The creation of these passes was data driven: we examined how coreference occurs in clinical records and created the different passes to resolve the different types of coreference that we observed.

Downloaded from jamia.bmj.com on August 4, 2012 - Published by group.bmj.com

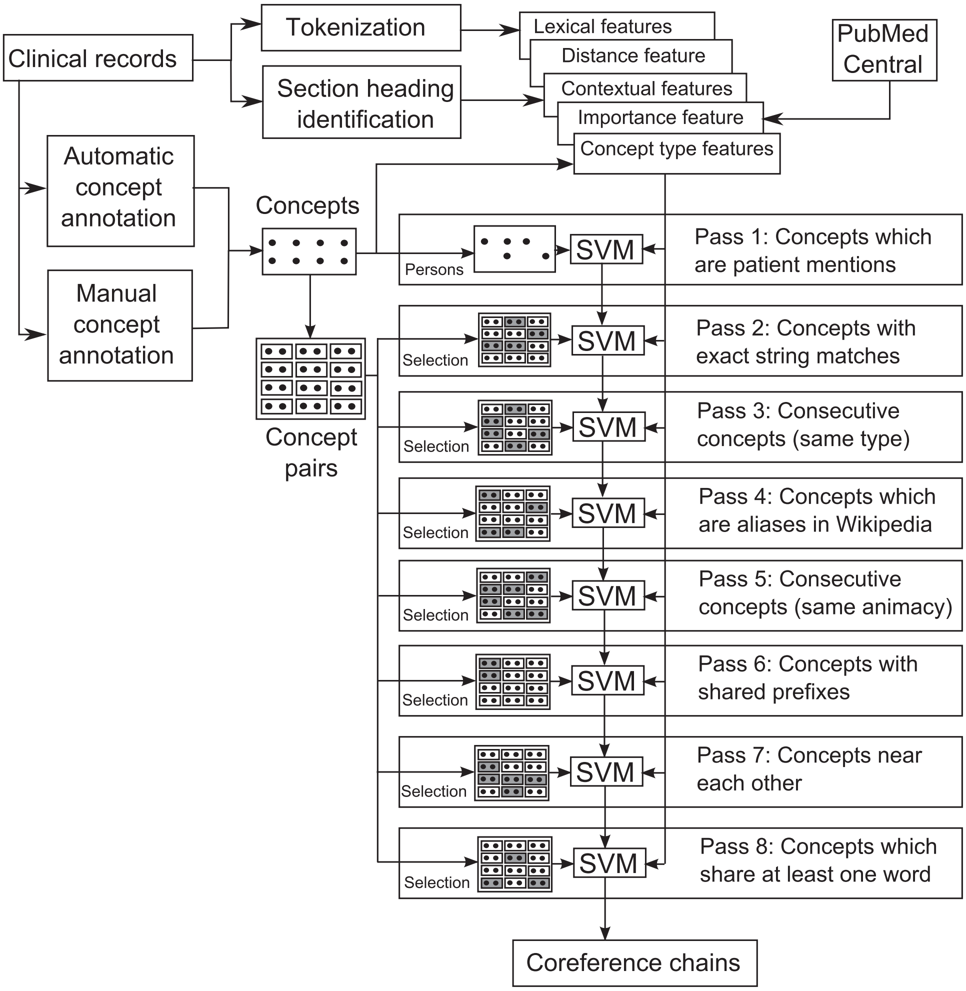
Clinical records
Tokenization
Section heading identification
Lexical features
Distance feature
Contextual features
Importance feature
Concept type features
PubMed Central
Automatic concept annotation
Manual concept annotation
Concepts
Concept pairs
Persons
Selection
SVM
Pass 1: Concepts which are patient mentions
Pass 2: Concepts with exact string matches
Pass 3: Consecutive concepts (same type)
Pass 4: Concepts which are aliases in Wikipedia
Pass 5: Consecutive concepts (same animacy)
Pass 6: Concepts with shared prefixes
Pass 7: Concepts near each other
Pass 8: Concepts which share at least one word
Coreference chains

Figure 1 Architecture for the multiple pass sieve strategy. SVM, support vector machine.

Pass 1: identification of patient mentions

The first model is unique because it does not operate in the same manner as the other models. In this pass, we utilize a classifier which identifies concepts referring to the patient. These mentions include 'the patient,' 'he,' 'she,' 'the infant,' etc. We train an SVM classifier to identify whether each concept refers to the patient. The manually annotated concepts and coreference chains provided for the patient records in the training data do not identify whether each concept refers to the patient. Therefore, we make an assumption that the largest coreference chain involving *people* (i2b2) or *person* (ODIE) concepts is the chain for references to the patient. This assumption appears to be true among a number of documents which we have inspected. An exception to this is the pathology reports in the ODIE dataset, which we discuss further in the Results section. All concepts belonging to the longest chain can then be used as positive training instances for a classifier, with the remaining people/person concepts being used as negative instances. The trained classifier is used to identify all concepts which are mentions of the patient in the testing data. Those concepts are then combined into a single coreference chain.

The classifier uses several features: the full text of the concept, individual tokens from the concept, the three tokens before/after the concept, character trigrams from the concept, the section header, and individual tokens from the section header. For the purposes of identifying sections within the clinical record, we consider any line ending in a colon to be a section header rather than more sophisticated techniques.[26 27]

While this initial pass operates on individual concepts, the remaining passes operate on pairs of concepts and identify whether each pair is coreferential or not.

Pass 2: coreference resolution between concepts with the same text

Selection criteria: concepts which share the same text

This pass will only consider pairs of concepts whose texts are the same. A loose definition of 'same' is used, where case is ignored, and Porter stemming is performed. Also, initial determiners and possessive pronouns are ignored. Under this relaxation, for instance, the concepts 'Propofol drips' and 'his propofol drip' would be considered the same. While previous coreference approaches have considered exact string matches to be coreferential,[15] we found that training a classifier to filter some pairs of concepts improved performance. The features used by the classifier are described in the next section.

Base set of features for passes 2 through 8

Passes 2 through 8 share a common set of features, although some of the passes have extra features used only by that pass. Each feature describes some property of a pair of concepts. Using features, a classifier will make a determination about whether those two concepts are coreferential or not. The features are detailed in table 1. Feature F3 enables the classifier to determine whether the concepts share the same concept type. Features F4 and F5 can be particularly useful for pronouns such as 'which' or 'This,' which tend to be coreferential with immediately adjacent

Downloaded from jamia.bmj.com on August 4, 2012 - Published by group.bmj.com

Table 1 Set of features used by passes 2 through 8

Feature name	Definition	Examples
F1	Concept type of the first concept	Person; Treatment; Test; etc
F2	Concept type of the second concept	Pronoun; Person; Problem; etc
F3	Concatenation of F1 and F2	Person-Test; Problem-Problem
F4	Full string of the first concept	Small cell lung cancer
F5	Full string of the second concept	Chemotherapy
F6	Individual tokens of the first concept	Small, cell, lung, cancer
F7	Individual tokens of the second concept	Chemotherapy
F8	All tokens found between the two concepts	Has, felt, cloudy, of, late, which, relates, to
F9	Number of tokens between the concepts	8
F10	Tokens which are found in both concepts	Lung, cancer
F11	Number of tokens found in both concepts	2
F12	Log term frequency of any shared tokens	17, 23
F13	Are the last tokens of both concepts the same?	True

concepts. F6 and F7 capture individual tokens in the concepts. Hence, the properties of coreference the classifier learns about 'small lung cancer' can also be applied to 'cervical cancer' because they both have the token 'cancer.' For example, both concepts have a high affinity for linking with mentions of 'tumors.'

Feature F8 provides information about the context of the two concepts. If there are more than 10 tokens between the concepts, this feature returns an empty set to avoid confusing the classifier with many non-contextual tokens. Feature F9 buckets the number of tokens between the concepts among: {1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 20, 50, 100}, with other distances being assigned to the closest bucket. Bucketing the distances in this way provides the classifier with more examples per feature value.

We have also included three features related to words which are shared between the two concepts. For instance, 'small cell lung cancer' and 'lung cancer' share the words 'lung' and 'cancer.' One feature (F10) indicates all of the shared words. Another feature (F11) indicates the number of shared words, which would be two in this case.

Sharing a common word such as 'the' or 'and' is less important for coreference detection than a shared word such as 'lung.' This importance is approximated by (F12) the log term frequency of the shared word in the 2011 PubMed Central corpus of scholarly biomedical articles, rounded to the lowest integer. For instance, the term 'lung' has a log term frequency of 17, while 'the' has a log term frequency of 25, and a rare term such as 'catecholamine' has a log term frequency of 11. Smaller values indicate less frequent terms, and therefore words which are more likely to be related.

Pass 3: identification of coreference between consecutive concepts of the same concept type

Selection criteria: consecutive pairs of concepts of the same concept type

Each concept is paired with the next concept in the text having the same concept type. In addition, pairs of concepts where at least one concept is a pronoun or was detected as a patient mention are disregarded. For example, in the following sentence several concepts have been marked:

> [This]$_{\text{pronoun}}$ is a 55-year-old male with [critical aortic stenosis]$_{\text{problem}}$ who was referred to [Dr John Doe]$_{\text{person}}$ for discussion for [surgical options]$_{\text{treatment}}$ to treat [this condition]$_{\text{problem}}$.

Pass 3 would extract the following pair of concepts for consideration:

> ([critical aortic stenosis]$_{\text{problem}}$, [this condition]$_{\text{problem}}$).

This pass implements the assumption that neighboring concepts of the same type are more likely to be coreferential.

Pass 4: resolving coreference when concepts are Wikipedia aliases

Selection criteria: any pairs of concepts that are mapped to the same article in Wikipedia

Wikipedia articles can have many aliases (redirects) to account for alternative spellings, misspellings, and synonymous terms. For example, the terms 'Hepatitis C' and 'Hep C' are both aliases within Wikipedia for the article about hepatitis C. Therefore, if a record contains both terms, the pair of concepts will be considered for coreference resolution in this pass. Other examples caught by this pass include 'A fib' and 'Atrial fibrillation,' 'Vtach' and 'Ventricular tachycardia,' as well as 'COPD' and 'Chronic obstructive pulmonary disease.'

This pass includes a feature indicating the canonical title of the Wikipedia article associated with the pair of concepts (eg, hepatitis C, atrial fibrillation). We also include a feature indicating the number of tokens in that Wikipedia article title. The intuition behind this feature is that longer titles are more likely to be correctly mapped.

Pass 5: recognition of coreferential consecutive concepts of the same animacy

Selection criteria: pairs of concepts that are either both animate or both inanimate

The concepts must also be consecutive (allowing for other concepts between them of the opposite animacy). The animacy of a concept was determined solely on the basis of concept type, where types 'person' and 'people' were marked as animate and all other concepts were marked as inanimate. Pairs of concepts involving at least one pronoun were ignored in this pass.

Pass 6: identification of coreference between concepts with shared prefixes

Selection criteria: pairs of concepts which have a common prefix of at least five characters

We assume that concepts which start with the same characters (a very simple type of word stem) are more likely to be the same entity, and likewise coreferential. Only concepts with no intervening concepts of the same prefix are paired. One such example

Table 2 Features selected for the automatic concept extraction method

Features with automatic i2b2 concepts (in order selected)	Features without automatic i2b2 concepts (in order selected)
Previous word stem	Previous word stem
Pattern-based entity	Pattern-based entity
i2b2 concept IOB type	1-token POS context
Uncased previous word	4-character suffix
Section name	2-character suffix
3-character suffix	Section name
Previous part of speech	Next GENIA phrase chunk
Current word	Uncased previous word
	5-character suffix
	Previous word
	1-character suffix

Downloaded from jamia.bmj.com on August 4, 2012 - Published by group.bmj.com

Table 3 Statistics about the training and testing data

	Training			Testing			
	Recs	Cons	Chains	Recs	Cons	Chains	Record types
i2b2							
Beth	115	24 392	2496	79	15 793	1816	Discharge
Partners	136	17 144	1792	94	11 713	1395	Discharge
Pitt	241	24 808	2765	151	16 361	2016	Discharge, Progress
i2b2 totals	492	66 344	7053	324	43 867	5227	
ODIE							
Mayo	118	1960	307	39	1515	208	Clinical, Pathology
Pitt	39	2319	302	27	1487	422	Discharge, Surgical Path., Other, Radiology
ODIE totals	157	4279	609	66	3002	630	
Totals	589	70 623	7662	410	46 869	5857	

Chains, coreference chains; Cons, concepts; Mayo, Mayo Clinic; Pitt, University of Pittsburgh Medical Center; Recs, records.

is the concepts 'hypotensive' and 'hypotension.' Stemming algorithms may not stem these two words identically. However, the UMLS SPECIALIST Lexicon is capable of enumerating morphological derivatives, including 'hypotensive/hypotension' and would be a better resource for identifying related concepts than relying on prefixes. The main benefits of an approach based on prefixes are speed and simplicity. In future work, we plan to integrate the UMLS SPECIALIST Lexicon to gain further improvements in performance.

This pass uses two features beyond the base set of features: the sets of tokens present in the section header for the first and second concept, respectively.

Pass 7: resolution of coreference between nearby concepts

Selection criteria: consecutive pairs of concepts which have at most seven tokens between them

This pass relies on the intuition that mentions which are close to each other are more likely to be coreferential, however no restriction is made on concept type, unlike in pass 3. One additional feature is added in this pass representing all word bigrams found between the mentions. A word bigram consists of two adjacent words. This pass was added when we noticed many phrases indicative of coreference such as '[concept], which,' where 'which' and '[concept]' are coreferential. Another example would be 'This [concept]' where 'This' and '[concept]' are coreferential. Rather than writing individual rules for every similar case, the machine learning classifier is able to learn the relevant patterns which indicate coreference.

Pass 8: identification of coreferential concepts which share at least one word

Selection criteria: pairs of concepts which have at least one word in common between them

This is a relaxation of pass 2 which requires the entire strings of both mentions to be the same. The shared word must not be a stopword or a single character. In addition, all pairs of concepts which meet the criteria for the second pass are not considered in this pass.

Training the classifiers

The classifiers used for all passes are trained independently of each other. All pairs of concepts from the training data meeting the selection criteria for a pass are used as training instances. Two concepts will be considered coreferential if any of the passes determines they are coreferential. During testing, after all passes have been executed, coreference chains are formed from all concepts which can be considered coreferential, using the transitive closure over pairs of concepts.

Automatic extraction of concepts

In order to determine the feasibility of using our coreference algorithm on completely unseen data, we ran a series of experiments with automatically annotated concepts. We utilized a pre-existing concept identification system,[28] which, given training data, automatically chooses the best set of features for concept identification. The system was trained on the ODIE concept data under two separate configurations: (1) automatically annotated concepts based on the i2b2 concept types (PROBLEM, TREATMENT, TEST, PERSON) were available to the ODIE concept classifier (which uses nine concept types), and (2) no automatic annotations were available to the ODIE concept classifier. The automatic concepts were represented in an IOB-style feature. The features selected for each configuration are shown in table 2. For more details on the individual features or feature selection process, see Roberts and Harabagiu.[28]

RESULTS

We evaluate our coreference approach by training the classifiers on the training portion of data made available during the i2b2 2011 Challenge, and by evaluating on the testing portion of data provided by that challenge. Both training and testing datasets were divided into two subsets, one which had been annotated under the ODIE standard, and another which had been annotated under the i2b2 standard. Table 3 summarizes the data. The records came from four hospital systems: Beth, Partners, Mayo

Table 4 Evaluation results for tasks 1B and 1C

		B^3			MUC			BLANC			CEAF			
Training set	Test	P	R	F	P	R	F	P	R	F	P	R	F	Avg
ODIE	ODIE	0.922	0.875	0.898	0.807	0.864	0.835	0.884	0.933	0.907	0.658	0.758	0.704	0.812
ODIE and i2b2	ODIE	0.921	0.878	0.899	0.808	0.874	0.840	0.926	0.926	0.926	0.661	0.771	0.712	0.817
i2b2	i2b2	0.963	0.946	0.954	0.834	0.887	0.860	0.961	0.975	0.968	0.873	0.907	0.890	0.901
ODIE and i2b2	i2b2	0.964	0.947	0.955	0.835	0.887	0.861	0.959	0.976	0.968	0.874	0.908	0.890	0.902

F, F score; P, precision; R, recall. Avg is an unweighted average of F scores for B^3, MUC, and CEAF.

Downloaded from jamia.bmj.com on August 4, 2012 - Published by group.bmj.com

Table 5 Evaluation of coreference results by concept type

ODIE concept type	F score	i2b2 concept type	F score
Anatomical site	0.733		
People	0.793	Person	0.902
Disease or syndrome	0.768	Problem	0.858
Sign or symptom	0.825		
Organ or tissue function	0.735		
Laboratory test or result	0.751	Test	0.823
Procedure	0.781	Treatment	0.828
Other	0.640		
None	0.658		
		Pronoun	0.665

The two annotation standards do not have a one-to-one mapping between concept types.

Table 6 Performance of coreference resolution on subsets of the clinical records, by type of record

Dataset	Record type	Number of records	Concepts in chains	Chains	Concepts per chain	F score
i2b2	Discharge	250	22 905	4314	5.3	0.904
	Progress	72	4933	913	5.4	0.935
ODIE	Discharge	6	390	55	7.1	0.948
	Pathology	20	254	61	4.2	0.817
	Clinical	19	1112	147	7.6	0.940
	Radiology	7	120	26	4.6	0.948
	Other	6	737	92	8.0	0.928
	Surgical pathology	8	193	38	5.1	0.945

Clinic, and the University of Pittsburgh Medical Center. No records were annotated using both i2b2 and ODIE guidelines. The ODIE subsets of the data were considerably smaller than the i2b2 subsets.

We performed several experiments to evaluate our method on these corpora. In the first experiment, we evaluated how our method performed on only the i2b2 portion of the data. Likewise, we experimented on only the ODIE portion of the data. Finally, we performed an experiment to evaluate our method when trained on both portions of the data. The i2b2 2011 Challenge used four official scoring metrics: B^3,[29] MUC,[30] BLANC,[31] and CEAF.[32] The challenge also included an official overall score which was the unweighted average of B^3, MUC, and CEAF, referred to in the tables below as Avg.

Table 4 shows the results of our evaluation. The scores on the smaller ODIE corpus are significantly lower than the scores on the i2b2 corpus. In both cases, training on all available data shows a small improvement. We analyzed the results by both concept types (table 5) and record types (table 6) in an effort to determine the reason for the lower scores on the ODIE data. Table 5 shows that F scores are lower for ODIE data across all concept types. This indicates that a difference in concept types was likely not the cause of the performance discrepancies between the i2b2 and ODIE datasets.

Furthermore, table 6 shows the performance of the automatic coreference resolution approach broken down by record type. The performance on the ODIE dataset is comparable to the i2b2 set on all record types except pathology reports. However, these constitute almost a third of the records and thus bring the overall score down significantly. The pathology reports are characteristically different from the other report types. For example, shown below are the chains for a single pathology report:

1. Invasive, grade 3 (of 4) adenocarcinoma arising from the tubular adenoma || Neoplasm
2. Rectal polyp base || The separately submitted polyp base || adenomatous mucosa
3. Colon || rectum || rectal || submucosa || a cauterized margin.

The first notable difference in these chains from many of the other types of reports is the absence of patient mentions. In discharge and progress notes, the patient is mentioned many times and constitutes a very large fraction of the mentions in coreference chains. Detecting patient mentions is also relatively easy in comparison to detecting other types of coreference. This leads to higher coreference scores for records which mention the patient frequently. The sixth column of table 6 shows the average number of concepts in coreference chains within each type of record. As expected, pathology records have the smallest coreference chains, largely due to the absence of patient mentions. An additional factor affecting the performance on pathology records is the semantic knowledge requirements for correctly detecting coreferential concepts. Each of the three example chains above contain terms which cannot be associated at the lexical level. One must know that a neoplasm (tumor) is caused by cancer and can therefore be used to refer to the cancer (adenocarcinoma) which caused it. These factors also explain why adding the i2b2 data did not significantly help performance when testing on ODIE data. The i2b2 data consist of only discharge and progress notes which do not contain nearly as many of the long, precise technical terms found in pathology notes.

Table 7 shows how well our method performs when only using some of the passes of the sieve method. The average shown in the last column increases as each pass is executed. The evaluation was performed using the i2b2 portion of the data. Pass 1, which identifies mentions of the patient and links them together, acts as a good baseline with a score only about 11 points lower than the score for all eight passes. This is reasonable because patient mentions usually form the largest coreference chain in each record. Also, other references are limited and

Table 7 Results obtained when running only a subset of the coreference passes using the i2b2 portion of the data

	B^3			MUC			BLANC			CEAF			
Passes	P	R	F	P	R	F	P	R	F	P	R	F	Avg
Pass 1	0.969	0.879	0.922	0.492	0.981	0.655	0.976	0.947	0.962	0.693	0.918	0.790	0.789
Passes 1−2	0.969	0.913	0.940	0.684	0.945	0.794	0.974	0.962	0.968	0.788	0.930	0.853	0.862
Passes 1−3	0.966	0.916	0.940	0.703	0.930	0.800	0.972	0.963	0.968	0.797	0.924	0.856	0.866
Passes 1−4	0.966	0.917	0.941	0.707	0.928	0.802	0.972	0.964	0.968	0.799	0.924	0.857	0.867
Passes 1−5	0.966	0.925	0.945	0.734	0.923	0.818	0.972	0.966	0.969	0.816	0.925	0.867	0.877
Passes 1−6	0.965	0.927	0.946	0.743	0.920	0.822	0.971	0.966	0.969	0.820	0.924	0.869	0.879
Passes 1−7	0.971	0.932	0.951	0.769	0.915	0.836	0.971	0.968	0.970	0.836	0.924	0.878	0.888
Passes 1−8	0.963	0.946	0.954	0.834	0.887	0.860	0.961	0.975	0.968	0.873	0.907	0.890	0.901

F, F score; P, precision; R, recall. Avg is an unweighted average of F scores for B^3, MUC, and CEAF.

Downloaded from jamia.bmj.com on August 4, 2012 - Published by group.bmj.com

Table 8 Results obtained when an individual pass is executed by itself

	B^3			MUC			BLANC			CEAF			
Passes	**P**	**R**	**F**	**P**	**R**	**F**	**P**	**R**	**F**	**P**	**R**	**F**	**Avg**
Pass 1	0.969	0.879	0.922	0.492	0.981	0.655	0.976	0.947	0.962	0.693	0.918	0.790	0.789
Pass 2	0.908	0.904	0.906	0.344	0.917	0.500	0.978	0.570	0.623	0.658	0.927	0.770	0.725
Pass 3	0.894	0.875	0.884	0.100	0.844	0.180	0.937	0.507	0.514	0.577	0.911	0.707	0.590
Pass 4	0.886	0.869	0.877	0.006	0.755	0.013	0.894	0.500	0.501	0.552	0.911	0.688	0.526
Pass 5	0.938	0.878	0.907	0.344	0.861	0.492	0.927	0.568	0.618	0.653	0.910	0.760	0.720
Pass 6	0.906	0.889	0.897	0.215	0.921	0.349	0.978	0.525	0.547	0.613	0.921	0.736	0.661
Pass 7	0.902	0.880	0.891	0.119	0.900	0.210	0.952	0.505	0.511	0.585	0.917	0.714	0.605
Pass 8	0.875	0.887	0.881	0.132	0.740	0.224	0.854	0.516	0.531	0.586	0.901	0.710	0.605

F, F score; P, precision; R, recall. Avg is an unweighted average of F scores for B^3, MUC, and CEAF.

usually only contain a few concepts. The second pass, which matches concepts that have the same string, has a very large impact, adding almost eight points to the average F score. Three other passes have an impact of almost a full point each. The first was pass 5, which incorporates information about consecutive concepts of the same animacy. The second was pass 7, which detects short patterns indicative of coreference between nearby concepts. Finally, pass 8 also has a large impact by identifying concepts which share words. Table 8 shows the results of the sieve method when only a single pass is executed. These results give a better idea of the strength of the passes on their own. Passes 1, 2, and 5 performed the strongest. The strong results from passes 2 and 5 are likely a result of their recall-oriented nature and the fact that these passes will link likely patient mentions as well.

Table 9 shows the performance of our end-to-end coreference approach on the ODIE test data, using automatically extracted concepts. Two methodologies were used: with and without i2b2-style automatic concepts. It appears that the performance was improved slightly by providing i2b2 concepts, despite the fact that i2b2 and ODIE used different concept types. It is unclear why the exact matching metric is higher for the second case than the partial matching.

DISCUSSION

Our approach achieved encouraging results while at the same time being simple and using mostly lexical features. During our analysis of the medical records and the occurrence of coreference within them, we observed that the majority of the references were of two kinds: (1) references to the patient and (2) nominal coreference. The majority of pronominal references other than simple construction such as 'which' and 'that,' were references to the patient. That is why we chose an approach which first tries to identify all mentions of the patient. Once those mentions have been identified, a large portion of the remaining references can be identified using fairly simple lexical features.

Despite the fact that our approach did not incorporate much syntactic, semantic, or pragmatic information, it is quite likely that such information would lead to even better results. Furthermore, information from existing medical ontologies such as UMLS[33] can be incorporated through the addition of a new pass which links concepts that are semantically similar. The limited use of external resources should enable the application of our approach in other domains with minimal reconfiguration. The various passes are not domain specific, relying primarily on vicinity and lexical features. Even pass 1 which detects patient mentions can be applied to other domains in which a single entity is mentioned predominantly in a document.

The order in which we performed the various passes roughly tries to fit the rule that more precise passes should be performed first, as suggested by Raghunathan *et al*.[15] One reason for this is that our method does not pass attributes about entities across passes as they do, therefore the order of the passes is not nearly as important. It is possible that further experiments regarding the ordering of passes could lead to additional gains in performance. A limitation to the extension of this method is the fact that passes are all trained independently and pairs of mentions linked together in all passes are combined together. The result is that adding any new passes cannot break the coreference chains being produced by earlier passes, the chains can only be made longer. This limitation could be remedied by incorporating passes which break chains, or through the use of a scoring-based approach such as Best-Link instead.

CONCLUSION

We were able to achieve promising results on the task of resolving coreference between concepts in medical records using a simple approach based on a multi-pass sieve which included machine learning classifiers in every pass. While our approach owes its inspiration to an existing method reported by Raghunathan *et al*,[15] we have adapted it in several important ways. The first is the inclusion of a classifier in each pass. The availability of a large corpus generously made available by i2b2/VA, University of Pittsburgh Medical Center and several other institutions allowed for a hybrid approach incorporating machine learning to outperform a purely rule-based approach.

Table 9 End to end results on the ODIE test dataset when using automatic concept recognition

		B^3			MUC			BLANC			CEAF			
Training set	**Metric**	**P**	**R**	**F**	**P**	**R**	**F**	**P**	**R**	**F**	**P**	**R**	**F**	**Avg**
ODIE	Exact	0.981	0.781	0.870	0.552	0.909	0.687	0.929	0.911	0.920	0.455	0.800	0.580	0.712
	Partial	0.965	0.780	0.863	0.552	0.860	0.672	0.922	0.911	0.916	0.450	0.797	0.575	0.713
ODIE and i2b2	Exact	0.978	0.782	0.869	0.561	0.910	0.694	0.931	0.913	0.921	0.456	0.799	0.581	0.715
	Partial	0.962	0.782	0.863	0.559	0.866	0.680	0.923	0.913	0.918	0.451	0.799	0.576	0.706

F, F score; P, precision; R, recall. Avg is an unweighted average of F scores for B^3, MUC, and CEAF.

Downloaded from jamia.bmj.com on August 4, 2012 - Published by group.bmj.com

Another significant diversion from the existing approach is that our method makes coreference decisions at the level of pairs of concepts, rather than finding a single antecedent for every mention. The information used by our method is primarily lexical. However, information about alternative spellings and synonyms of concepts from Wikipedia was also incorporated. Exploration of the addition of even more semantic (UMLS SPECIALIST Lexicon, SNOMED CT,[34] and distributional similarity techniques[35]) along with pragmatic information[36] will be the goal of our immediate future work due to the encouraging results obtained by this approach.

Acknowledgments We would like to acknowledge the efforts of the organizers for the 2011 i2b2 Challenge, who made this work possible.

Funding The 2011 i2b2/VA challenge and the workshop are funded in part by grant number 2U54LM008748 on Informatics for Integrating Biology to the Bedside from the National Library of Medicine. This challenge and workshop are also supported by the 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 VA Informatics and Computing Infrastructure (VINCI), VA HSR HIR 08-204, and the National Institutes of Health, National Library of Medicine under grant number R13LM010743-01. MedQuist, the largest transcription technology and services vendor, co-sponsored the 2011 i2b2/VA challenge meeting at AMIA.

Competing interests None.

Provenance and peer review Not commissioned; externally peer reviewed.

REFERENCES

1. **Uzuner O,** South BR, Shen S, *et al*. 2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text. *J Am Med Inform Assoc* 2011;**18**:552—6.
2. **Hobbs JR.** Resolving pronoun references. *Lingua* 1978;**44**:339—52.
3. **Ng V.** Semantic class induction and coreference resolution. *Annual Meeting-Association for Computational Linguistics*; Prague, Czech Republic. Stroudsburg, PA: Association for Computational Linguistics, 2007;**45**.
4. **Ng V,** Cardie C. Improving machine learning approaches to coreference resolution. *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*; University of Pennsylvania, USA. Stroudsburg, PA: Association for Computational Linguistics, 2002.
5. **Baldwin B.** CogNIAC: high precision coreference with limited knowledge and linguistic resources. *Proceedings of a Workshop on Operational Factors in Practical, Robust Anaphora Resolution for Unrestricted Texts*. Stroudsburg, PA: Association for Computational Linguistics, 1997.
6. **Yang X,** Su J. Coreference resolution using semantic Relatedness information from automatically Discovered patterns. *Annual Meeting-Association for Computational Linguistics*; Prague, Czech Republic. Stroudsburg, PA: Association for Computational Linguistics, 2007;**45**.
7. **Stoyanov V,** Cardie C, Gilbert N, *et al*. Coreference resolution with reconcile. *Proceedings of the ACL 2010 Conference Short Papers*; Uppsala, Sweden. Stroudsburg, PA: Association for Computational Linguistics, 2010.
8. **Recasens M,** Marquez L, Sapena E, *et al*. SemEval-2010 Task 1: coreference resolution in multiple languages. *Proceedings of the 5th International Workshop on Semantic Evaluation*; Uppsala, Sweden. Stroudsburg, PA: Association for Computational Linguistics, 2010.
9. **Bengston E,** Roth D. Understanding the value of features for coreference resolution. *Proceedings of the Conference on Empirical Methods in Natural Language Processing*; Waikiki, Honolulu, Hawaii. Stroudsburg, PA: Association for Computational Linguistics, 2008.
10. **Soon WM,** Ng HT, Lim DCY. *A machine learning approach to coreference resolution of noun phrases. Computational linguistics*. Cambridge, MA: MIT Press, 2001;**27**.
11. **Versley Y,** Ponzetto SP, Poesio M, *et al*. BART: a modular toolkit for coreference resolution. *Proceedings HLT-Demonstrations '08 Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies: Demo Session*; Columbus, OH. Stroudsburg, PA: Association for Computational Linguistics, 2008.
12. **Haghighi A,** Klein D. Coreference resolution in a modular, entity-centered model. *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational linguistics*; Los Angeles, CA. Stroudsburg, PA: Association for Computational Linguistics, 2010.
13. **Ng V.** Unsupervised models for coreference resolution. *EMNLP '08 Proceedings of the Conference on Empirical Methods in Natural Language Processing*; Waikiki, Honolulu, Hawaii. Stroudsburg, PA: Association for Computational Linguistics, 2008.
14. **Poon H,** Domingos P. Joint unsupervised coreference resolution with Markov logic. *Proceedings of the Conference on Empirical Methods in Natural Language Processing*; Waikiki, Honolulu, Hawaii. Stroudsburg, PA: Association for Computational Linguistics, 2008:650—9.
15. **Raghunathan K,** Lee H, Rangarajan S, *et al*. *A multi-pass sieve for coreference resolution, Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*; MIT Stata Center, MA. Stroudsburg, PA: Association for Computational Linguistics, 2010.
16. **Nicolae C,** Nicolae G. BestCut: a graph algorithm for coreference resolution. *EMNLP '06 Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*; Sydney, Australia. Stroudsburg, PA: Association for Computational Linguistics, 2006.
17. **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.
18. **Wang Y,** Melton GB, Pakhomov S. It's about this and that: a description of anaphoric expressions in clinical text. *AMIA Annu Symp Proc* 2011;**2011**:1471—80.
19. **He TY,** Uzuner O, Szolovits P. Coreference resolution on entities and events for hospital discharge summaries. Thesis (M. Eng.), Massachusetts Institute of Technology, 2007
20. **Zheng J,** Chapman WW, Miller TA, *et al*. A system for coreference resolution for the clinical narrative. *J Am Med Inform Assoc*. Published Online First. doi:10.1136/amiajnl-2011-000599
21. **Segura-Bedmar I,** Crespo M, de Pablo-Sánchez C, *et al*. Resolving anaphoras for the extraction of drug-drug interactions in pharmacological documents. *BMC Bioinformatics* 2010;**11**(Suppl 2):S1.
22. **Vlachos A,** Gasperin C, Lewin I, *et al*. Bootstrapping the recognition and anaphoric linking of named entities in Drosophila articles. *Pac Symp Biocomput* 2006:100—11.
23. **Pustejovsky J,** Castaño J, Zhang J, *et al*. Robust relational parsing over biomedical literature: extracting inhibit relations. *Pac Symp Biocomput* 2002:362—73.
24. **Uzuner O,** Forbush T, Shen S, et al. *i2b2/VA 2011 Co-reference Annotation Guidelines for the Clinical Domain*. 2011. https://www.i2b2.org/NLP/Coreference/assets/CoreferenceGuidelines.pdf
25. **Fan RE,** Chang KW, Hsieh CJ, *et al*. LIBLINEAR: a library for large linear classification. *J Machine Learn Res* 2008;**9**.
26. **Denny JC,** Spickard A 3rd, Johnson KB, *et al*. Evaluation of a method to identify and categorize section headers in clinical documents. *J Am Med Inform Assoc* 2009;**16**:806—15.
27. **Li Y,** Gorman SL, Elhadad N. Section classification in clinical notes using supervised hidden Markov model. *IHI '10 Proceedings of the 1st ACM International Health Informatics Symposium*, Arlington, VA. New York: ACM, 2010.
28. **Roberts K,** Harabagiu SM. A flexible framework for deriving assertions from electronic medical records. *J Am Med Inform Assoc* 2011;**18**:568—73.
29. **Amit B,** Baldwin B. Algorithms for scoring coreference chains. *Proceedings of the First International Conference on Language Resources and Evaluation Workshop on Linguistic Coreference*; Granada, Spain. Paris, France: European Language Resources Association (ELRA), 1998.
30. **Vilain M,** Burger J, Aberdeen J, *et al*. A model-theoretic coreference scoring scheme. *Proceedings of the 6th Message Understanding Conference (MUC6)*; Columbia, MD. Stroudsburg, PA: Association for Computational Linguistics, 1995.
31. **Recasens M,** Hovy E. BLANC: Implementing the Rand index for coreference evaluation. *Natural Language Engineering*. 2010.
32. **Luo X.** On coreference resolution performance metrics. *HLT '05 Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*; Vancouver, BC, Canada. Stroudsburg, PA: Association for Computational Linguistics, 2005.
33. **Aronson AR.** Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. *Proceedings of the AMIA Symposium*. 2001;**17**.
34. **Stearns MQ,** Price C, Spackman KA, *et al*. SNOMED clinical terms: overview of the development process and project status. *Proceedings of the AMIA Symposium*; Washington, DC. Bethesda, MD: American Medical Informatics Association, 2001.
35. **Lee L.** Measures of distributional similarity. *Proceedings of the 37th annual meeting of ACL*; College Park, Maryland, USA. Stroudsburg, PA: Association for Computational Linguistics, 1999.
36. **Iida R,** Inui K, Takamura H, *et al*. Incorporating contextual cues in trainable models for coreference resolution. *Proceedings of the 10th EACAL Workshop on the Computational Treatment of Anaphora*; Budapest, Hungary. Stroudsburg, PA: Association for Computational Linguistics, 2003.

Downloaded from jamia.bmj.com on August 4, 2012 - Published by group.bmj.com

A supervised framework for resolving coreference in clinical records

Bryan Rink, Kirk Roberts and Sanda M Harabagiu

J Am Med Inform Assoc published online May 19, 2012
doi: 10.1136/amiajnl-2012-000810

Updated information and services can be found at:
http://jamia.bmj.com/content/early/2012/05/19/amiajnl-2012-000810.full.html

These include:

References This article cites 7 articles, 3 of which can be accessed free at:
http://jamia.bmj.com/content/early/2012/05/19/amiajnl-2012-000810.full.html#ref-list-1

P<P Published online May 19, 2012 in advance of the print journal.

Email alerting service Receive free email alerts when new articles cite this article. Sign up in the box at the top right corner of the online article.

Notes

Advance online articles have been peer reviewed, accepted for publication, edited and typeset, but have not not yet appeared in the paper journal. Advance online articles are citable and establish publication priority; they are indexed by PubMed from initial publication. Citations to Advance online articles must include the digital object identifier (DOIs) and date of initial publication.

To request permissions go to:
http://group.bmj.com/group/rights-licensing/permissions

To order reprints go to:
http://journals.bmj.com/cgi/reprintform

To subscribe to BMJ go to:
http://group.bmj.com/subscribe/