A classification approach to coreference in discharge summaries: 2011 i2b2 challenge

Yan Xu,[1,2] Jiahua Liu,[2,3] Jiajun Wu,[2,4] Yue Wang,[2,5] Zhuowen Tu,[2,6,7] Jian-Tao Sun,[2] Junichi Tsujii,[2] Eric I-Chao Chang[2]

► Additional tables are published online only. To view these files please visit the journal online (http://dx.doi.org/10.1136/amiajnl-2011-000734).

[1]State Key Laboratory of Software Development Environment, Key Laboratory of Biomechanics and Mechanobiology of Ministry of Education, Beihang University, Beijing, China
[2]Microsoft Research Asia, Beijing, China
[3]Department of Computer Science and Technology, Tsinghua University, Beijing, China
[4]Institute for Interdisciplinary Information Sciences, Tsinghua University, Beijing, China
[5]School of Information Security Engineering, Shanghai Jiaotong University, Shanghai, China
[6]Department of Neurology, Lab of Neuro Imaging, University of California, Los Angeles, USA
[7]Department of Computer Science, Lab of Neuro Imaging, University of California, Los Angeles, USA

Correspondence to
Dr Eric Chang, Microsoft Research Asia, T2-14463, No 5 Danling Street, Haidian District, Beijing 100080, PR China; eric.chang@microsoft.com

Received 1 December 2011
Accepted 19 March 2012
Published Online First 13 April 2012

ABSTRACT

Objective To create a highly accurate coreference system in discharge summaries for the 2011 i2b2 challenge. The coreference categories include Person, Problem, Treatment, and Test.

Design An integrated coreference resolution system was developed by exploiting Person attributes, contextual semantic clues, and world knowledge. It includes three subsystems: Person coreference system based on three Person attributes, Problem/Treatment/Test system based on numerous contextual semantic extractors and world knowledge, and Pronoun system based on a multi-class support vector machine classifier. The three Person attributes are patient, relative and hospital personnel. Contextual semantic extractors include anatomy, position, medication, indicator, temporal, spatial, section, modifier, equipment, operation, and assertion. The world knowledge is extracted from external resources such as Wikipedia.

Measurements Micro-averaged precision, recall and F-measure in MUC, BCubed and CEAF were used to evaluate results.

Results The system achieved an overall micro-averaged precision, recall and F-measure of 0.906, 0.925, and 0.915, respectively, on test data (from four hospitals) released by the challenge organizers. It achieved a precision, recall and F-measure of 0.905, 0.920 and 0.913, respectively, on test data without Pittsburgh data. We ranked the first out of 20 competing teams. Among the four sub-tasks on Person, Problem, Treatment, and Test, the highest F-measure was seen for Person coreference.

Conclusions This system achieved encouraging results. The Person system can determine whether personal pronouns and proper names are coreferent or not. The Problem/Treatment/Test system benefits from both world knowledge in evaluating the similarity of two mentions and contextual semantic extractors in identifying semantic clues. The Pronoun system can automatically detect whether a Pronoun mention is coreferent to that of the other four types. This study demonstrates that it is feasible to accomplish the coreference task in discharge summaries.

INTRODUCTION

Coreference is defined as two mentions referring to the same entity in a sentence or document.[1] A mention is typically a named entity. In this i2b2 challenge, participants are given unstructured text of a discharge summary with the mentions. A mention is in the form *c="<mention string>" line:start_token line:end_token||e="<category>"*, where category is denoted as *<person|problem|treatment|test|pronoun>*. Our task is to resolve the coreference problem of each category to construct a chain of mentions which refer to the same entity. All chains including singletons recognized by a system are used for evaluation. There are five types of mentions: Person, Problem, Treatment, Test, and Pronoun. A coreference chain belongs to one of the four semantic classes (ie, Person, Problem, Treatment, and Test).[2] A Pronoun can be coreferent to a mention of the other four types. In Person, many phrases or words are used to describe the patient in a discharge summary, such as "the patient", "pt", "who", "she", "her", "Mary", "Mr. Kotefooks, Dasha", "you", and "your." In Problem, very different surface expressions such as "laceration on the right edge" and "tongue biting" describe the same Problem. On the other hand, two occurrences of "fever" in "He returns from the nursing home with fever, leukocytosis, and azotemia" and "On 6-12-91 the patient was admitted to Ingtermst.gay Health Center with fever, hypertension, and diarrhea" are not considered to denote the same entities, because they occur in different episodes. Coreference chains of Treatment and Test have difficulties similar to those of Problem. That is, the same entities can be described by very different surface expressions, while the same surface expressions denote different entities in different episodes. Pronoun mentions are restricted to 15 different surface expressions such as "this", "that", "which", "it", etc. Unlike Problem/Treatment/Test, these surface expressions do not contain any characteristic strings which can be used for mention chain detection for the other types. Supplementary table 1 (available at http://jamia.bmj.com) lists examples of the five types of coreference. The texts highlighted with bold font are coreferent pairs.

The training set and the test set are provided by the i2b2 organizers[2] and are collected from four hospitals. The training set includes 492 labeled discharge summaries with 1597 chains for Person, 2630 chains for Problem, 1895 chains for Treatment and 891 chains for Test. The test set consists of 322 discharge summaries. The test set includes 173 discharge summaries from Partners Healthcare and Beth Israel Deaconess Medical Center and 149 discharge summaries from University of Pittsburgh hospital Medical Center. Unstructured discharge summaries and all the mentions in them with their corresponding semantic types are provided for both the training set and the test set. In addition, all the coreference chains are provided for the training set.

The key observation we made on the training set is that different types of coreference have very different characteristics, which require different

coreference resolvers. As the examples in supplementary table 1 (available at http://jamia.bmj.com) show, the mention strings in coreference chains of Problem/Treatment/Test often share the same substrings (eg, <Serratia urospesis, sepsis>, <An echocardiogram, Echocardiogram>, etc). On the other hand, a coreference chain of Person contains many mentions in the form of pronoun, for which common substring features are of no use. For Problem/Treatment/Test coreference, a wide variety of synonymous expressions such as abbreviations, jargons, and aliases are also used to denote the same entities, which are the main causes of difficulties in coreference resolution. We need to use external resources such as Wikipedia to gather such synonyms of the same concepts. Such diverse synonyms do not exist for Person.

Furthermore, a single large coreference chain (ie, the specific patient of the summary) normally exists among chains of Person, while no such single dominant chains exist for Problem/Treatment/Test. As the previous example of "fever" shows, we have to distinguish occurrences with the same concept in different episodes. We do not distinguish occurrences of the same Person in different spatio-temporal contexts (ie, different episodes).

Taking these different characteristics of the types into consideration, we decided to build three separate resolvers, one for Person, one for Problem/Treatment/Test and one for Pronoun. They have different architecture and use different sets of features.

In Person coreference, since the distinction of patient and non-patient is crucial, we constructed a binary classifier for this distinction. All mentions of person are partitioned by three mutually exclusive attributes: patient, relative, and hospital personnel. Coreference only occurs between mentions with the same attribute. In Problem/Treatment/Test coreference resolvers, to treat synonyms and distinguish different episodes is crucial. We exploited rich world knowledge to gather various synonyms of the same concepts. In addition, we explored a broad range of contextual semantic extractors to find semantic clues by which different spatio-temporal contexts and subtle difference of semantics are distinguished. The extractors include those for anatomy, position, medication, indicator, temporal, spatial, section, modifier, equipment, operation, and assertion.

In this paper, we present the whole coreference system that was built for the 2011 coreference task, which consists of three coreference resolvers: (1) Person coreference resolver, which uses the output of a binary classifier that learns whether a Person mention refers to a patient or not; (2) Problem/Treatment/Test coreference resolver, which combines the world knowledge to treat synonyms with various contextual semantic extractors, by which the resolver distinguishes different entities of the same concepts in different spatio-temporal contexts or subtle differences of semantics; and (3) Pronoun coreference based on a multi-class support vector machine (SVM) classifier. The experimental results demonstrate that the approach is appropriate for the coreference task in discharge summaries.

RELATED WORK

Coreference resolution has been an active area of research for over 15 years.[3] Three important classes of coreference models have been developed—namely, the mention-pair model,[4] the entity-mention model,[5] and the ranking model.[6]

The mention-pair model aims at a classifier which judges whether a given pair of two noun phrases (NPs) is coreferential or not. Several different strategies have been proposed to create a training set of positive and negative pairs. The simplest method is to take all the C_n^2 pairs of mentions in a training text.[7] This method would produce too many negative pairs which might bias the trained classifier. To reduce the bias caused by negative pairs, some work used only negative pairs with NPs which intervene between a positive pair.[8] Other work[9 10] filtered some pairs to reduce the number of training instances. Because the i2b2 challenge data contain the types of mentions and except for the Pronoun type, only mentions in the same semantic types can be coreferential, we can easily filter irrelevant negative pairs and a preliminary experiment showed that, by filtering them, we could avoid the problem of negative bias. Therefore, we used the simplest method for generating C_n^2 pairs for training by filtering out pairs of different semantic types.

The second major decision is how to judge coreferential relations among all pairs of NPs. The common methods are to use classification resolver,[11] cluster resolver,[12] or classification and cluster resolver.[13] A classification resolver can be based on decision trees, rule learners, memory-based learners, or statistical learners (eg, maximum entropy models, voted perceptrons and SVM). A cluster resolver can be the closest-first clustering, best-first clustering, correlation clustering, graph partitioning algorithm, or Dempster–Shafer rule.

The entity-mention model defines coreference resolution as a clustering problem, instead of a pairwise classification. It resolves whether the current NP is related to a preceding cluster or not. There are three types of features, which are all relevance, most relevance, and any relevance between the NP and the preceding clusters. The ranking mechanism tries to select the anaphoric NP which has the highest rank with the candidate antecedent. These models have been developed for text in the general domain, in which several dominant mention chains of the same semantic type co-exist and only syntactic (but poor semantic) clues such as pronouns, definite noun phrases exist for them. However, coreferences in discharge summaries are very different in nature from those of the general domain: (1) there is only one dominant large chain (the patient); (2) semantic types of mentions are restricted and give strong clues, etc. Instead of difficulties in resolving competing dominant chains of the same semantic types, we have to deal with different sets of difficulties such as (1) the same concepts but different entities in different contexts have to be distinguished; (2) many synonyms which lack syntactic cues such as definite articles have to be recognized, etc. Therefore, we use the simplest resolver framework of a mention classification as the basic framework and focus on enriching features.

There is some related work on feature design. Features are classified into two sets: internal feature set and external feature set. The internal feature set consists of string-matching features, syntactic features, grammatical features, and semantic features. The external feature set is extracted from Wikipedia,[14] Freebase,[15] WordNet,[16] and Yago.[17]

In the medical science domain, there has been little existing work on coreference. In He's work,[18] a semantic coreference resolver was proposed for medical practitioners, treatments, diseases, symptoms, and medical tests. Their feature sets consist of orthographic, semantic, lexical features, syntactic and morphological features, temporal features, and miscellaneous. The resources are UMLS,[19] and C4.5 decision tree.

METHODS

In this section, we describe in detail our method for the coreference challenge. Our overall approach comprises the following steps (see figure 1): preprocessing, creating positive/negative instances of mention-pairs, training classification systems using SVM classifiers, generating the SVM confidence scores for each instance, selecting pairs as coreference by best-first clustering,

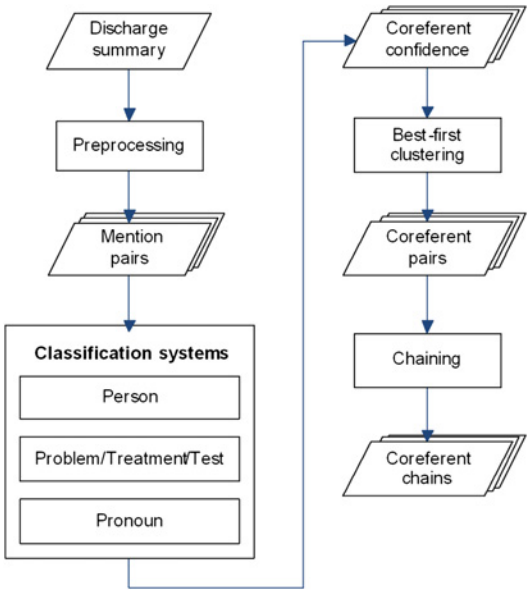

Figure 1 The overall approach.

forming coreferent pairs according to clustering results, and combining all coreferent pairs to produce chains. In the following sections, we will describe the details of each step.

Preprocessing

This preprocessing is only used for world knowledge features and string match features. Some mentions are too specific owing to various modifiers, such as "her CT scan" and "a CT scan." We remove modifiers on the left/right of the mention; if the mention contains a preposition, the preposition and the content after it are also removed. After the preprocessing step, the above two mentions both become "CT scan."

Mention pair

Mention pair[8] is used to resolve a coreference. Given *n* mentions, we consider each of the C(n, 2) mention pairs <i, j> and decide whether i and j are coreferent.

Classification methods

In this task, the mention types are Person, Problem, Treatment, Test and Pronoun. Each coreference relation belongs to one of these types except for Pronoun. The most salient characteristic of Person mentions is a wide range of personal pronouns, possessive pronouns, and reflexive pronouns. The coreference between proper name and pronoun in general is difficult because a pronoun gives only very limited information such as Singular/Plural, First/Second/Third person, etc. Furthermore, since mentions of more than one person appear in a single article, it is a great challenge to identify chains among them correctly. However, if we confine ourselves to a discharge summary, and if a Person mention is judged to be a patient mention, it almost always belongs to the single, largest coreference chain of the patient. It is rare in a discharge summary that more than one patient appears. Therefore, to classify a Person mention into patient/non-patient is of utmost importance in the discharge summary domain.

For Problem/Treatment/Test coreference, though the same types of medical events may occur several times, they may not be coreferential but constitute entities in different spatio-temporal contexts or subtle difference of semantics. Construction of correct chains of these types requires richer semantic clues in a local context which distinguishes spatio-temporal contexts or subtle differences of semantics.

As for the resolver for Pronoun, the crucial step is to identify the semantic type (Person, Problem, Treatment or Test). Once the semantic type is decided, we choose the closest mention of the same type as its antecedent. Though simple, the method works very well. This shows that coreference resolution in discharge summaries is a very different problem from that of the general domain. Figure 2 shows the flow diagram of the overall classification system. In Person coreference system and the Problem/Treatment/Test coreference system, we use pair-mentions as instances, whereas in the Pronoun coreference system, mentions are regarded as instances. In person system, the patient classifier is adopted and the corresponding classification result is used as a Person feature.

Person coreference

As we stated above, Person coreference systems for the general domain cannot be directly applied to discharge summaries. Systems trained for coreference resolution in newswire corpus assume that quite a few different people or groups of people appear in a single article and they play equally significant roles in the article. In contrast, people mentioned in a discharge

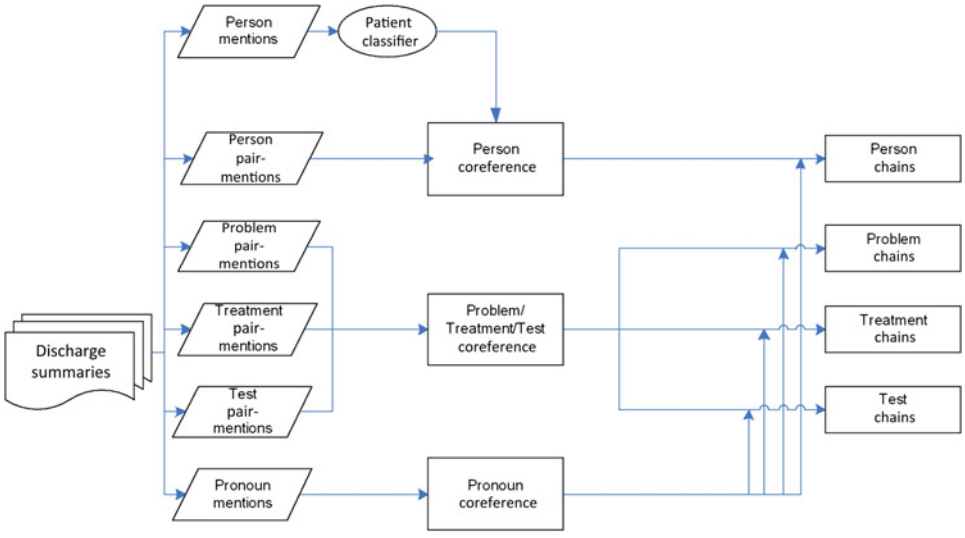

Figure 2 The overall classification method.

summary (or a medical record, in general) are limited to three classes: patient, relatives, and hospital personnel. To obtain such information on the class of Person mentions (including pronoun and proper name), we introduce a feature which we called Patient-class (explained below). The Person feature set is in table 1. The machine learning algorithm we used was binary-SVM.[20]

Patient-class feature

A binary SVM classifier is trained to distinguish the patient from the rest of the mention classes (relatives and medical personnel) in Person mentions. Assuming that there is only one patient in a record, we assign all patient mentions to a long single chain. It is relatively easy to pick out all these patient mentions by a few keywords from training chains. To train the classifier, we use all the mentions from the patient chain as positive instances, while negative instances are the remaining Person mentions. The feature set is given in table 2. The classifier outcome is used to create the "Patient-class feature" in table 1. The high performance (F-measure: 0.996) of this binary classifier makes the feature very reliable.

Problem/Treatment/Test coreference

The types of Problem, Treatment and Test are classes specific to the domain of medical records. In this domain, various phrases can express the same concepts. Identifying these synonymous phrases can reduce false negatives and improve recall. In order to find synonymous phrases other than those that appear in the training data, we leveraged existing resources on world knowledge such as Wikipedia. On the other hand, many mentions are not coreferential even if they have nearly identical literal strings because the contexts are different—for example, different episodes of the same medication. Distinguishing these mentions can reduce false positives and improve the precision. We designed 11 contextual semantic extractors to provide contextual information and semantic clues for distinguishing different semantic contexts.

In the following subsection, first we introduce the resources on world knowledge we used and contextual semantic extractors. Then we present the Problem/Treatment/Test coreference resolver based on a binary classifier using mention pairs as instances.

World knowledge

Existing resources on world knowledge can provide a set of surface expressions with the same or synonymous concepts. Coreference can benefit from world knowledge. In our work, we used the external resources on world knowledge such as Wikipedia, WordNet, and MSRA. Below we will briefly describe these three resources.

Wikipedia[14] is a free encyclopedia database. The redirected link, the phrase based on bold font and the anchor information from Wikipedia are used to find synonyms, aliases, and abbreviations.

WordNet[16] is used to seek the synonyms and aliases of words in mentions.

MSRA resource[21–23] is a knowledge base consisting of mentions, instances, attributes and values, and relationships. It can provide synonyms and also various expressions of the same phrase. For example, other expressions of "vancomycin" are topical vancomycin, antibiotic vancomycin, dry van, vanguard 550, los van, and van. Probase is described further at http://research.microsoft.com/en-us/projects/probase/.

Contextual semantic extractors

Mentions of the types of Problem/Treatment/Test heavily have to be distinguished according to the semantic context. Although we distinguish the same concepts in different spatio-temporal contexts, actual clues which signal different contexts are diverse. For example, two mentions of the same Problem which appear in different anatomical regions should not be coreferential. *Pain* in the head is not coreferential to *pain* in the leg. If the two mentions of the same medication appear in different modes such as orally or intravenously, they are not

Table 1 Features for Person coreference

Feature and feature perspectives	Possible value	Description
Patient-class feature	0, 1, 2	Neither is patient (0), both patients (1), others (2)
Distance between sentences	0, 1, 2, 3, …	
Distance between mentions	0, 1, 2, 3, …	
String match	0, 1	The same string, false (0), true (1)
Levenshtein distance between similarity	(0, 1)	Levenshtein distance between two mentions
Number	0, 1, 2	Both singular or plural, false (0), true (1), unknown (2)
Gender	0, 1, 2	Both the same genders, false (0), true (1), unknown (2)
Apposition	0, 1	Is appositive, false (0), true (1)
Alias	0, 1	Two mentions are abbreviations or acronyms, false (0), true (1)
Who	0, 1	The previous mention, false (0), true (1),
Name match	0, 1	Removing stop words (dr, dr., mr, mrs, ms, md, m.d., m.d, ",", m, m., :), using substring matching, false (0), true (1)
Relative match	0, 1	Both the same relatives, false (0), true (1)
Department match	0, 1	Both the same departments, false (0), true (1)
Doctor title match	0, 1	Both the same doctor titles, false (0), true (1)
Doctor general match	0, 1	Matching general doctor, false (0), true (1)
Twin/triplet	0, 1	Both the same twins/triplets, false (0), true (1)
We	0, 1	Both 'we' information, false (0), true (1)
You	0, 1	Both 'you' information, false (0), true (1)
I	0, 1	Both 'I' information, false (0), true (1)
Pronoun match	0, 1	Matching the previous mention, false (0), true (1)

Table 2 Features for the binary classifier of patient/non-patient

Feature and feature perspectives	Possible value	Description
Semantic		
Key word of patient	0, 1	Mr, mr., mrs, mrs., ms, ms., miss, yo-, y.o., y/o, year-old, …
Key word of doctor	0, 1	Dr, dr., md, m.d., m.d, …
Key word of doctor title	0, 1	Anesthesiologist, orthodontist, dentist, …
Key word of department	0, 1	Anesthesiology, electrophysiology, …
Key word of general department	0, 1	Team, service
Key word of general doctor	0, 1	Doctor, pcp, author, dict, attend, provider
Key word of relative	0, 1	Wife, brother, sister, sibling, nephew, …
Name	0, 1	A full uppercase mention, the first letter capital for each word in a mention
Last *n* line doctor	0, 1	Doctor's name in the last *n* lines
Twin or triplet information	0, 1	Baby 1, 2, 3, 4, …
Preceded by non-patient	0, 1	Dr XXX, he, …
Signed information	0, 1	Signed, dictator
Previous sentence		
Next sentence		
Grammatical		
Pronouns we	0, 1	Our, we, us, ourselves
Pronouns I	0, 1	I, me, my, myself
Pronouns you	0, 1	You, your, yourself
Pronouns they	0, 1	They, their, them, themselves
Pronouns he/she most	0, 1	<He, him, his, himself>, <She, her, hers, herself> the most number of pronoun chosen
Who	0, 1	The previous mention
Appositive	0, 1	
Lexicon		
First one/two/three words before mention		
First one/two/three words after mention		
First one/two/three words in between mention, problem		
Last one/two/three words in between mention, problem		
First one/two/three words in between mention, treatment		
Last one/two/three words in between mention, treatment		
First one/two/three words in between mention, test		
Last one/two/three words in between mention, test		

coreferential. If two same Test mentions have different values, they are not coreferential, either. Therefore, we developed a set of contextual semantic extractors whose outputs contribute to recognition of different spatio-temporal contexts and subtle differences of semantics.

Anatomy extractor

Two mentions of the same symptom are not coreferential when they appear in different anatomical regions. For example, "The patient continued to suffer from edema of the left upper extremity and a vascular radiology consult revealed **a thrombosis of the left subclavian vein** extending into the axillary vein." and "There was **some thrombosis of the left internal jugular vein** as well." Although the two mentions share the same string of "**thrombosis**", "**subclavian**" and "**jugular**" it denotes different anatomical parts. To recognize different anatomical parts in mentions is particularly useful when mentions are almost indistinguishable by string matching. To explore anatomy information, we developed an anatomy extractor by a dictionary search. The dictionary with a tree of anatomical concepts was constructed from UMLS,[19] SNOMED,[24] MESH,[25] and RadLex.[26]

Position extractor

The same concepts of Problem/Treatment /Test can appear in different positions. In the following two sentences, "The **right upper** extremity was positive for *abrasions.*" and "The **right lower** extremity was positive for *abrasions*." the bold font indicates positions and the italic font indicates mentions. The two different positions can be used to distinguish the two mentions with a same string. A dictionary search is used to extract semantic clues of positions. This dictionary of positions was generated from the training data, which may not be comprehensive.

Medication information extractor

This extractor deals with drugs, modes, dosages, frequencies, durations, reasons, and "List" or "Narrative." Please refer to Uzuner *et al*[27] for more details. When drugs, modes, dosages, frequencies, or durations are different, the two mentions are not coreferent in Treatment coreference. The dictionaries and regular expressions are collected from medication data provided by i2b2 organizers.[27]

Indicator extractor

In Test mentions, there are various indicators, such as "wbc", "rbc", "hgb", and "hct." When the indicator values are different, the two mentions are not coreferent. Such as "**WBC**—5.7 RBC—3.10 * Hgb—8.9 *" and "BLOOD **WBC**—6.2 RBC—3.10 * Hgb—8.9 * Hct—26.2." The indicator dictionary is collected from training data. We developed a dictionary search algorithm to hunt for indicators.

Temporal extractor

Temporal information is an important semantic clue. For Treatment coreference, various operations and drugs at different times are independent. For Test, examinations with the same name but taken at different times are not coreferent. In the task, the temporal information is separated into two parts, one is explicit dates; the other is inferred dates. Explicit dates are obtained by some regular expressions. The inferred dates are obtained by some key dates and section information, such as "admission date", "evaluation date", "operation date", "transfer date", and "discharge date."

Spatial extractor

Spatial information is a useful semantic clue only for Treatment. For two literally identical medication mentions, when one appears in emergency department and the other appears in an intensive care unit, they are not coreferent. In spatial extractor, spatial information is approximately divided into four categories: home, transfer hospital, inpatient hospital, and individual department. Chunks of information and key words are used to extract the location.

Section mapping engine

Medical records include some sections, such as "history of present illness", "past medical history", and "medications on admission."[28] They are a rich source of clinical information and semantic clues. Let us take an example. A mention of *CT scan* in the section of "history of present illness" and another mention in the "physical examination" section are independent of each other, although they can be exactly matched with string-matching. We implement the algorithm[27] for the section extractor.

Modifier extractor

Whether two mentions are coreferent is affected by some specific modifiers for Test. The modifier dictionary is compiled from training data, such as "initial", "recent", and "prior."

Equipment extractor

Equipment is also a semantic clue since medical tests are often named by the equipment. Equipment dictionary consists of UMLS words with suffix "-graphy" "-gram" "-metry" or "-scopy", and RadLex words under "imaging modality" class.

Operation extractor

Many of the Treatment mentions are surgical operations, which can be recognised as a semantic clue for the Treatment. Operation dictionary consists of UMLS words with the suffix "-tomy" and "-plasty."

Assertion extractor

Assertions of Problem mentions are from the i2b2 2010 task, including "not associated with the patient", "hypothetical", "conditional", "possible", "absent", and "present."[29] It is an important clue to resolve ambiguities of Problem mentions. We directly use the information from i2b2 2010 task to search for assertions of problems.

Of these extractors, position and section are features for Problem, Treatment and Test; temporal for Treatment and Test; anatomy for Problem and Test; assertion for Problem; medication, spatial and operation for Treatment; indicator, modifier and equipment for Test (see supplementary table 2 available at http://jamia.bmj.com).

Feature sets and model training

The feature set is given in table 3. The machine learning algorithm is binary-SVM.

Pronoun coreference

The Pronoun category consists of 15 pronouns, of which *this*, *that*, *which* and *it* appear frequently and comprise the biggest portion. Each pronoun is either an independent mention, or coreferential to a previous mention of one of the four semantic types. Aiming at the aforementioned characteristic, we used a multi-class SVM classifier[30] to determine whether it is or not coreferential that paired with the most adjacent Person, Treatment or Test mention. The feature set is listed in supplementary table 3 (available at http://jamia.bmj.com).

RESULTS

We submitted three systems according to various SVM thresholds (see supplementary table 4 available at http://jamia.bmj.com). The performance was evaluated using three measures: MUC, B-CUBED, and CEAF.[31] The results were micro-averaged precision (P), recall (R) and F-measure (F). The final precision (P) is the average of three micro-averaged precisions produced by each evaluation metric; the final recall (R) and F-measure (F) are computed similarly (equations 1, 2, 3).

$$\text{Precision}(P) = (P_{MUC} + P_{B-CUBED} + P_{CEAF})/3 \quad (1)$$

$$\text{Recall}(\text{R}) = (R_{MUC} + R_{B-CUBED} + R_{CEAF})/3 \quad (2)$$

$$\text{F—}measure(P) = (F_{MUC} + F_{B-CUBED} + F_{CEAF})/3 \quad (3)$$

Table 4 summarizes the performance of our submitted three systems for coreference on two groups of test data. The two groups are 322 from all hospitals and 173 without Pittsburgh data. Our systems achieved the first place in this task. The performance of our best system, system 3, yielded an F-measure of 0.915, a precision of 0.906, and a recall of 0.925 using all test data; and an F-measure of 0.913, a precision of 0.905, and a recall of 0.920 using all test data except the Pittsburgh data.

Supplementary table 5 (available at http://jamia.bmj.com) shows the performance of patient classification based on 10-fold cross-validation experiments. The training data of 494 discharge summaries included 13 694 positive instances and 5605 negative instances. The proper names and pronouns of patient belonged to positive instances. The algorithm achieved a promising F-measure of 99.6%.

Supplementary table 6 (available at http://jamia.bmj.com) shows the performance of the Person coreference system with and without patient learned features. The performance of the system with the Patient-class feature was significantly improved by 23.1% compared with the system without the Patient-class feature. The experiment demonstrated that the Patient-Class feature is helpful in improving the Person coreference system.

Supplementary table 7 (available at http://jamia.bmj.com) summarizes contributions of world knowledge for the Problem/Treatment/Test system. The baseline system used all the features except those from world knowledge. The experiment showed that world knowledge is helpful in improving the system performance. Public domain resources included Wikipedia and WordNet. As shown in supplementary table 7 (available at http://jamia.bmj.com), the improvement of the Problem coreference was the highest in three coreference types. The features of open sources and MSRA resource had the same contributions of 0.2%.

Supplementary table 8 (available at http://jamia.bmj.com) summarizes contributions of various contextual semantic

Table 3 Features for Problem/Treatment/Test coreference

Feature and feature perspectives	Possible value	Description
World knowledge		
Wiki page match	0, 1	Mentions redirected to the same page
Wiki bold name match	0, 1	
Wiki anchor match	0, 1	
WordNet match	0, 1	
MSRA resource match	0, 1	
Contextual semantic extractors		
Anatomy extractor	0, 1, 2	Both the same anatomical structure false (0), true (1), unknown (2)
Position extractor	0, 1,2	Both the same position, false (0), true (1), unknown (2)
Indicator extractor	0, 1, 2	Both the same indicator value, false (0), true (1), unknown (2)
Temporal extractor	0, 1, 2	Both the same time, false (0), true (1), unknown (2)
Spatial extractor	0, 1, 2	Both the same space, false (0), true (1), unknown (2)
Section extractor	1, 2, …, n^2	i and j may belong to n possible sections
Modifier extractor	0, 1, 2	Both the same modifier, false (0), true (1), unknown (2)
Equipment extractor	0, 1, 2	Both the same equipment, false (0), true (1), unknown (2)
Operation extractor	0, 1, 2	Both the same operation, false (0), true (1), unknown (2)
Assertion extractor	1, 2, …, 6^2	I and j may have six possible assertions
Medication extractors		
Drug	0, 1	Both the same drug, false (0), true (1)
Mode	0, 1, 2, …, 29	29 Categories or unknown
Dosage	0, 1	Both the same dosage, false (0), true (1)
Frequency	0, 1	Both the same frequency, false (0), true (1)
Duration	0, 1	Both the same duration, false (0), true (1)
"List" or "Narrative"	0, 1	Both the same "List" or "Narrative", false (0), true (1)
Time of first mention	0, 1, 2, 3	Past (0), present (1), future (2), unknown (3)
Time of second mention	0, 1, 2, 3	Past (0), present (1), future (2), unknown (3)
Episode of first mention	0, 1, 2, 3	Start (0), continue (1), stop (2), unknown (3)
Episode of second mention	0, 1, 2, 3	Start (0), continue (1), stop (2), unknown (3)
Condition of first mention	0, 1, 2, 3	Factual (0), suggestion (1), conditional (2), unknown (3)
Condition of second mention	0, 1, 2, 3	Factual (0), suggestion (1), conditional (2), unknown (3)
Distance		
Distance	0, 1, 2…	Sentence distance
Grammar		
Article	1, 2, …, 3^2	(a|an), (the|his|her|…), (NULL) between i and j
Orthographic		
Head noun match	0, 1	Both the same head noun, false (0), true (1)
Contains	0, 1	I contains j or j contains i, false (0), true (1)
Capital match	0, 1	Both the same initials, false (0), true (1)
Substring match	0, 1	Both the same substring, false (0), true (1)
Cos distance	(0, 1)	
Semantic		
Word match		Cartesian product of the tokens in i and j
Procedure match	0, 1	False (0), true (1)

extractors for the Problem/Treatment/Test system. The baseline system used all the features except those from contextual semantic extractors. The experiment showed that the section extractor and the assertion extractor had the most significant contribution by 0.4% in the whole F-measure. Assertion, medication, and equipment had major contributions for Problem, Treatment, and Test by 2.2%, 1%, and 2% improvement.

While supplementary table 8 (available online at http://jamia.bmj.com) shows cumulative improvements by adding features, supplementary table 9 (available at http://jamia.bmj.com) shows contributions of a single feature by adding each of them to the baseline. XX/YY in Increase in supplementary table 9 (available online at http://jamia.bmj.com) means that XX and YY are the increments which the feature brings to the baseline system and the cumulative system respectively. Supplementary table 9 (available at http://jamia.bmj.com) shows that Medication, Indicator, Spatial, Modifier and Equipment are independent—that is, the increment by the single feature is more or less maintained in the cumulative system. The independent contribution by the Section feature is the largest among the features.

DISCUSSION

Person

Our approach can identify most of the mentions of the patient. Mistakes occur when a woman is giving birth to one or more children: the system cannot determine if the patient is the mother, the infant, or in some cases, one of the newborn twins or triplets. This is because "mother" is mistaken for a relative, and it is difficult to distinguish between multiple infants. Another unsolved problem arises when coreferring "I" to one of the physicians. In some records, "I" refers to the attending doctor; in other records it refers to the dictator. More information is needed to determine the author of the record.

Table 4 Micro-averaged results for coreference in discharge summaries

	BCubed			Muc			CEAF			Ave		
	P	R	F	P	R	F	P	R	F	P	R	F
All test data												
System 1												
All	0.978	0.96	0.969	0.847	0.906	0.875	0.883	0.92	0.901	0.903	0.929	0.915
Test	0.98	0.96	0.97	0.346	0.62	0.444	0.922	0.958	0.94	0.749	0.846	0.785
Person	0.982	0.951	0.966	0.968	0.986	0.977	0.887	0.927	0.906	0.946	0.955	0.950
Problem	0.965	0.949	0.957	0.708	0.79	0.746	0.889	0.91	0.899	0.854	0.883	0.867
Treatment	0.958	0.937	0.947	0.673	0.741	0.705	0.858	0.875	0.866	0.830	0.851	0.839
System 2												
All	0.976	0.962	0.969	0.854	0.898	0.876	0.887	0.916	0.901	0.906	0.925	0.915
Test	0.978	0.96	0.969	0.366	0.62	0.46	0.922	0.958	0.94	0.755	0.846	0.790
Person	0.982	0.946	0.964	0.964	0.985	0.975	0.873	0.923	0.897	0.940	0.951	0.945
Problem	0.964	0.949	0.956	0.715	0.79	0.751	0.889	0.91	0.9	0.856	0.883	0.869
Treatment	0.949	0.945	0.947	0.721	0.719	0.72	0.873	0.86	0.866	0.848	0.841	0.844
System 3												
All	0.975	0.962	0.968	0.856	0.897	0.876	0.888	0.915	0.902	**0.906**	**0.925**	**0.915**
Test	0.977	0.962	0.969	0.406	0.615	0.489	0.927	0.957	0.942	0.770	0.845	0.800
Person	0.982	0.946	0.964	0.964	0.985	0.975	0.873	0.923	0.897	0.940	0.951	0.945
Problem	0.964	0.949	0.956	0.715	0.79	0.751	0.889	0.91	0.9	0.856	0.883	0.869
Treatment	0.949	0.945	0.947	0.721	0.719	0.72	0.873	0.86	0.866	0.848	0.841	0.844
All test data but Pittsburgh												
System 1												
All	0.977	0.962	0.969	0.839	0.894	0.866	0.885	0.917	0.901	0.900	0.924	0.912
Test	0.981	0.965	0.973	0.327	0.581	0.418	0.927	0.959	0.943	0.745	0.835	0.778
Person	0.98	0.959	0.969	0.973	0.984	0.978	0.906	0.922	0.914	0.953	0.955	0.954
Problem	0.964	0.947	0.955	0.712	0.794	0.751	0.881	0.906	0.893	0.852	0.882	0.866
Treatment	0.956	0.942	0.949	0.742	0.72	0.699	0.852	0.865	0.858	0.850	0.842	0.835
System 2												
All	0.974	0.964	0.969	0.848	0.885	0.866	0.89	0.913	0.901	0.904	0.921	0.912
Test	0.979	0.965	0.972	0.356	0.584	0.442	0.927	0.96	0.943	0.754	0.836	0.786
Person	0.98	0.951	0.965	0.967	0.983	0.975	0.883	0.915	0.898	0.943	0.950	0.946
Problem	0.962	0.948	0.955	0.721	0.794	0.756	0.882	0.906	0.894	0.855	0.883	0.868
Treatment	0.947	0.949	0.948	0.747	0.721	0.733	0.868	0.85	0.858	0.854	0.840	0.846
System 3												
All	0.974	0.964	0.969	0.85	0.884	0.867	0.891	0.912	0.902	**0.905**	**0.920**	**0.913**
Test	0.978	0.967	0.972	0.401	0.586	0.476	0.932	0.958	0.945	0.770	0.837	0.798
Person	0.98	0.951	0.965	0.967	0.983	0.975	0.883	0.915	0.898	0.943	0.950	0.946
Problem	0.962	0.948	0.955	0.721	0.794	0.756	0.882	0.906	0.894	0.855	0.883	0.868
Treatment	0.947	0.949	0.948	0.747	0.721	0.733	0.868	0.85	0.858	0.854	0.840	0.846

Problem

Identical disease and symptoms are not coreferential if they happen at different times. A "fever" can last for some time and be referred to in later descriptions, but multiple episodes of "fever" are not uncommon. There is no definite answer as to whether a medical problem is always chronic (long-lasting) or acute (of short duration). Another difficulty is in pairing two Problem mentions with no string overlap. For example, "a 3×3 cm mass" and "tumor on the left side" can be coreferential in a discharge summary. However, "tumor" and "mass" frequently occur in the same discharge summary, but most of them are singletons in annotated discharge summaries. This is because a patient has several masses and several tumors. To identify coreferential pairs among them requires more sophisticated context processing (eg, their sizes, body parts on which they are seen, the point of time at which they are identified, etc) than currently used. Another problem is lack of domain knowledge. Although "a posterior scalp laceration" and "the patient's head wound" may be identified as coreference in a summary, the current system does not have access to such comprehensive synonym pairs.

Treatment

Forty per cent of the error pairs in Treatment have identical strings but they are not coreferential. A typical example is when the same drug is applied at different times, dosage or by a different route of administration. In most records, drugs in "medication on admission" and "medication on discharge" are not coreferential; but they are coreferential in some other records. This is beyond the capability of our current classifier.

Test

The results of a medical test are usually in a table with header and contents. But in medical records, these tables are converted to consecutive lines whose columns are no longer aligned. In cases where some of the fields are left blank, reconstructing the table may be ambiguous and cause problems. This is a major source of error in our results for Test.

CONCLUSION

The paper describes a coreference resolver for discharge summaries which consist of three different subresolvers. Unlike coreference resolvers used for the general domain, these resolvers

are specifically designed to exploit the characteristics of coreferences in discharge summaries. The resolver for Person uses the Patient-class feature produced by a specially designed binary classifier for Patient, while the resolver for Problem/Treatment/Test exploits the results produced by a set of contextual semantic extractors and the large synonym dictionary constructed from the existing resources on world knowledge. The resolver for pronoun focuses on semantic type recognition of a pronoun mention. The three subsystems are built based on binary-SVM and multi-class SVM classifiers. We demonstrated that the attribution of Person in terms of patient versus non-patient (relatives and hospital personnel) improves the performance of the resolver for Person significantly. For the Problem/Treatment/Test resolver, the synonym dictionary constructed from world knowledge and a set of contextual semantic extractors improve both recall and precision. For Pronoun coreference, syntactic information accurately identifies the grammatical component of Pronoun. A comparison with the results of other methods indicates that the classification approach is promising in this challenge.

In the future, on the one hand, we will add domain knowledge (such as UMLS and MESH) to help recognize the same entities with different surface expressions such as "scalp laceration" and "head wound." Additionally, we will need to have a deeper understanding of context to treat cases such as the "mass"/"tumor" example. More sophisticated machine learning methods may be required.

Acknowledgments We would like to thank the organizers of the i2b2 NLP challenge.

Funding This work was supported by Microsoft Research Asia (MSR Asia). The work was also supported by MSRA eHealth grant, ONR N000140910099, NSF CAREER award IIS-0844566, Grant 61073077 from National Science Foundation of China and Grant SKLSDE-2011ZX-13 from State Key Laboratory of Software Development Environment in Beihang University in China.

Competing interests None.

Patient consent Obtained.

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

Provenance and peer review Not commissioned; externally peer reviewed.

REFERENCES

1. **Uzuner Ö,** Bodnari A, Shen S, *et al*. Evaluating the state of the art in coreference resolution for electronic medical records. *J Am Med Inform Assoc* 2012;**19**:786—91.
2. *The 2011 i2b2 Challenge*. https://www.i2b2.org/NLP/Coreference/Call.php
3. **Ng V.** Supervised noun phrase coreference research: the first fifteen years. *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (ACL 2010)*. Uppsala, Sweden. 2010:1396—411.
4. **Aone C,** Bennett SW. Evaluating automated and manual acquisition of anaphora resolution strategies. *Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics (ACL 1995)*. Cambridge, Massachusetts, USA. 1995:122—9.
5. **McCallum A,** Wellner B. Toward conditional models of identity uncertainty with application to proper noun coreference. *Proceedings of the 2003 International Joint Coreference on Artificial Intelligence (IJCAI 2003)*. Acapulco, Mexico. 2003.
6. **Connolly D,** Burger JD, Day DS. A machine learning approach to anaphoric reference. *Proceedings of the International Conference on New Methods in Language Processing (NeMLaP 1994)*. Manchester, UK. 1994:255—61.
7. **Soon WM,** Ng HT, Lim CY. Corpus-based learning for noun phrase coreference resolution. *Proceedings of Conference on Empirical Methods in Natural Language Processing (EMNLP 1999)*. College Park, MD, USA. 1999:285—91.
8. **Soon WM,** Ng HT, Lim DCY. A machine learning approach to coreference resolution of noun phrases. *Comput Ling* 2001;**27**:512—44.
9. **Ng V,** Cardie C. Identifying anaphoric and non-anaphoric noun phrases to improve coreference resolution. *Proceedings of the 19th International Conference on Computational Linguistics (COLING 2002)*. Taipei, Taiwan. 2002:730—6.
10. **Strube M,** Rapp S, Muller C. The influence of minimum edit distance on reference resolution. *Proceedings of Conference on Empirical Methods in Natural Language Processing (EMNLP 2002)*. Philadelphia, PA, USA. 2002:312—19.
11. **Daelemans W,** Bosch AVD. *Memory-based Language Processing*. Cambridge, UK: Cambridge University Press, 2005.
12. **Versley Y,** Moschitti A, Poesio M, *et al*. Coreference systems based on kernels methods. *Proceedings of the 22th International Conference on Computational Linguistics (COLING 2008)*. Manchester, UK. 2008:961—8.
13. **Rahman A,** Ng V. Supervised models for coreference resolution. *Proceedings of Conference on Empirical Methods in Natural Language Processing (EMNLP 2009)*. Singapore. 2009:968—77.
14. **Wikipedia.** http://www.wikipedia.org/
15. **Freebase.** http://www.freebase.com/
16. **WordNet.** http://wordnet.princeton.edu/
17. **Yago.** http://www.mpi-inf.mpg.de/yago-naga/yago/
18. **He TY.** *Coreference Resolution on Entities and Events for Hospital Discharge Summaries*. M.S. Thesis. Cambridge, MA, USA: MIT, 2007.
19. *UMLS Knowledge Base*. http://www.nlm.nih.gov/research/umls
20. **Schölkopf B,** Burges CJC, Smola AJ. *Advances in Kernel Methods—Support Vector Learning*. Cambridge, MA, USA: MIT Press, 1999.
21. **Song YQ,** Wang HX, Wang ZY, *et al*. Short text conceptualization using a probabilistic knowledgebase. *Proceedings of the 2011 International Joint Coreference on Artificial Intelligence (IJCAI 2011)*. Barcelona, Catalonia, Spain. 2011.
22. **Lee TS,** Wang ZY, Wang HX, *et al*. Web scale taxonomy cleansing. *Proceedings of the 37th International Conference on Very Large Data Bases (VIDB 2011)*. Seattle, WA, USA. 2011.
23. **Zhang F,** Shi SM, Liu J, *et al*. Nonlinear evidence fusion and propagation for hyponymy relation mining. *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics (ACL 2011)*. Portland, Oregon, USA. 2011.
24. *SNOMED Knowledge Base*. http://www.nlm.nih.gov/research/umls/Snomed/snomed_main.html
25. *MESH Knowledge Base*. http://www.ncbi.nlm.nih.gov/mesh
26. *RadLex Knowledge Base*. http://www.radlex.org/
27. **Uzuner O,** Solti I, Xia F, *et al*. Community annotation experiment for ground truth generation for the i2b2 medication challenge. *J Am Med Inform Assoc* 2010;**17**:519—23.
28. **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.
29. **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.
30. *Multi-class SVM*. http://svmlight.joachims.org/svm_multiclass.html
31. **Luo XQ.** On coreference resolution performance metrics. *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL 2005)*. Ann Arbor, Michigan, USA. 2005:25—32.