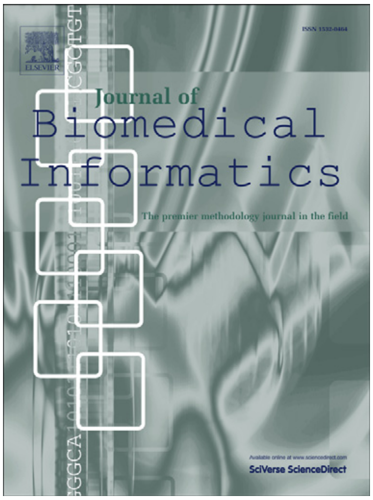
Lexical patterns, features and knowledge resources for coreference resolution in clinical notes

Phil Gooch, Abdul Roudsari



PII: S1532-0464(12)00042-1

DOI: 10.1016/j.jbi.2012.02.012

Reference: YJBIN 1892

To appear in: *Journal of Biomedical Informatics*

Received Date: 2 February 2012

Accepted Date: 29 February 2012

Lexical patterns, features and knowledge resources for coreference resolution in clinical notes

Phil Gooch[1], Abdul Roudsari[1,2]

[1]Centre for Health Informatics, City University, London, UK

[2]School of Health Information Science, University of Victoria, BC, Canada

Correspondence: Phil Gooch, Centre for Health Informatics, School of Informatics, City University, Northampton Square, London EC1V 0HB, UK; e-mail: Philip.Gooch.1@city.ac.uk, Phone (+44) 020 7040 3914, Mobile: (+44) (0)7547 139 793, Fax (+44) 020 7040 0244

Abstract

Generation of entity coreference chains provides a means to extract linked narrative events from clinical notes, but despite being a well-researched topic in natural language processing, general-purpose coreference tools perform poorly on clinical texts. This paper presents a knowledge-centric and pattern-based approach to resolving coreference across a wide variety of clinical records from two corpora (Ontology Development and Information Extraction (ODIE) and i2b2/VA), and describes a method for generating coreference chains using progressively pruned linked lists that reduces the search space and facilitates evaluation by a number of metrics. Independent evaluation results give an F-measure for each corpus of 79.2% and 87.5%, respectively. A baseline of blind coreference of mentions of the same class gives F-measures of 65.3% and 51.9% respectively. For the ODIE corpus, recall is significantly improved over the baseline ($p<0.05$) but overall there was no statistically significant improvement in F-measure ($p>0.05$). For the i2b2/VA corpus, recall, precision, and F-measure are significantly improved over the baseline ($p<0.05$). Overall, our approach offers performance at least as good as human annotators and greatly increased performance over general-purpose tools. The system uses a number of open-source components that are available to download.

Keywords: natural language processing, coreference resolution, knowledge engineering, clinical records, algorithms

1. Introduction

In linguistics, the relationship of coreference holds when two or more expressions or *mentions* (typically noun phrases) refer to the same external entity, independent of context or order within the text. The semantic relation between the expressions is one of identity. Coreference can be considered a specific type of anaphoric relation where a later expression (anaphor) has some semantic relation to an earlier expression (antecedent) and disambiguation of the anaphor is dependent on knowledge of the antecedent.[1][2] In a general anaphoric association, the semantic relation may be of identity, but not necessarily; for example, anaphor and antecedent may be in a part-whole relationship.

Pronominal coreference considers the resolution of pronouns back to their correct antecedents, while *bridging coreference* considers the resolution of definite descriptors (e.g. '*the procedure*') and semantically equivalent terms (e.g. synonyms and hypernyms) back to the specific antecedent. Relations may be both coreferent and anaphoric, for example '*initially the patient refused* **bronchoscopy** *but agreed to* **it** *later*', as the anaphor '*it*' can only be understood in relation to the antecedent '*bronchoscopy*', and both '*it*' and '*bronchoscopy*' refer to the same, external concept (a bronchoscopy procedure). With compound terms, the relationships can be multiple and more complex, for example:

the patient's *head* wound laceration … **her** *scalp* laceration

where there is potentially both an anaphoric and coreferent relationship between '*her*' and '*the patient*' (given the world knowledge that the patient is female), '*scalp*' is anaphoric to '*head*' in a meronym–holonym (part–whole) relationship, and '*head wound laceration*' and '*scalp laceration*' are potentially coreferent if they refer to the same injury.

Automated systems for coreference resolution have generally transitioned from rule-based heuristics to a variety of supervised machine learning approaches. Rule-based approaches have been dominated by research on pronominal coreference on general texts by Lappin, Leass and Mitkov (reviewed in Gasperin[1]), which typically involve backward-looking search from a given pronoun to the best antecedent. Antecedent ranking rules consider factors such as gender, number, token distance and sentence recency; syntax such as grammatical role (subject, direct object, indirect object), person, and position; and discourse models such as centering theory[3]. Supervised machine learning approaches have, until recently, been dominated by the *mention-pair* model, which treats coreference resolution as a binary classification problem between pairs of mentions, but has been criticized for considering each mention in isolation and not the wider context[4]. More recent models consider the task as a cluster-ranking problem that utilizes grammatical, syntactic, semantic and discourse-based contextual features[3][4].

2. Background

Resolution of coreference is particularly important in clinical notes, such as discharge summaries, as it is often required to uncover implicit and contextual information. For example

> **Patient** suffers from lower back pain. **He** takes *Vicoprofen* for this but *the medication* is not managing **his** discomfort.

To the human reader, it is clear that the patient’s lower back pain is managed unsuccessfully with Vicoprofen. Yet without a method for resolving ‘*he*’ and ‘*his*’ back to ‘*patient*’, ‘*this*’ and ‘*his discomfort*’ back to ‘*lower back pain*’, and ‘*the medication*’ back to ‘*Vicoprofen*’, there is no way that this can be inferred computationally. Moreover, this example shows the importance of resolution of complete *chains* of coreference (‘*patient–he–his*’, ‘*lower back pain–this–his discomfort, ‘Vicoprofen–the medication’*) in enabling this information to be extracted.

However, in a review of coreference methodologies, Zheng et al.[5] noted that there was a lack of both manually annotated corpora and automated systems for identifying coreference within the clinical domain. They concluded that an approach that identifies patterns specific to clinical texts, combined with adaptation of more general methods, would be a necessary first step towards a solution.[5] However, existing, general-purpose coreference tools, such as the BART Coreference Toolkit[6] or Stanford Deterministic Coreference Resolution System[7] – even when retrained for the clinical domain – perform poorly on clinical texts, where recall is particularly low, varying from 0-35%[8]. This is perhaps not surprising, as transcribed clinical notes present particular problems for identification of co-referring terms, such as

Spelling inconsistencies and errors.

Use of abbreviations without expansion and which may be ambiguous, e.g. ‘PT’ may abbreviate ‘patient’, ‘physiotherapy’, or ‘prothrombin time’.

Name anonymisation potentially resulting in the same personal name being replaced with a different string during deidentification, and the anonymised name may not match the patient’s gender

Potentially wide scope of resolution for personal pronouns. For example, ‘*he*’ might refer to ‘*the patient*’ mentioned several sentences or paragraphs previously, as intervening paragraphs may have discussed, for example, laboratory results.

Despite this, there are few evaluation reports of automated approaches to coreference resolution in clinical texts. Romauch[9] developed a knowledge-based system using the MetaMap Transfer (MMTx) application and the Unified Medical Language System (UMLS) to resolve definite descriptors in clinical practice guidelines, reporting an *F*-measure of 75.8%. Analysis revealed incorrect UMLS mappings made by MMTx, inadequate acronym/abbreviation detection, and incomplete coreference chains as sources of error. For hospital discharge summaries, He[10] used a supervised decision-tree classifier with a mention-pair model to resolve coreference chains of Person, Symptom, Disease, Medication, and Test mentions, and achieved a mean *F*-measure of 81.0% (ranging from 95.0% for Medications to 50.6% for Tests). Analysis revealed incomplete handling of temporal context, lack of knowledge-based handling of synonym and hypernym relationships, and lack of acronym/abbreviation detection, as the main factors affecting system recall.

Two manually annotated corpora of clinical anaphoric relations have recently been made available as part of the 2011 i2b2/VA challenge on coreference resolution[11]: the Ontology Development and Information Extraction (ODIE) corpus[12], consisting of de-identified clinical notes and pathology reports from the Mayo Clinic, and discharge summaries, progress notes, radiology reports, surgical pathology reports, and progress notes from the University of Pittsburgh Medical Center (UPMC), and the i2b2/VA corpus[13], consisting of de-identified discharge summaries from Partners HealthCare, Beth Israel Deaconess Medical Center, and UPMC. An evaluation script has also been released[14] so that systems can be measured against a number of published metrics, including B^3[15], MUC[16], CEAF[17] and BLANC[18]. Using these metrics, Zheng et al.[19] have very recently published results for a system that used a variety of supervised machine learning approaches to resolve coreference in the ODIE corpus. Using a support vector machine with a radial basis function, they achieved a mean *F*-measure (over all metrics) of 53.1%.

There is controversy over which is the most valid metric for evaluating coreference chains, particularly when dealing with coreference of system-generated mentions not in the key (gold standard) set, leniency in handling split coreference chains, and singletons (mentions with no coreferents). The reasons for this are beyond the scope of this paper (for a detailed discussion refer to Cai and Strube[20], and Zheng et al.[5]). Briefly, these metrics perform complex set-wise comparisons of coreference chains between the key set and the system output under evaluation. Under certain conditions, however, they can give unexpected results. For example, for a null system output (i.e. no coreference relations and no mentions) against an arbitrary key set containing ~44000 mentions and ~5200 coreference chains, *F*-measures of 0.936, 0.5, and 0.686 are reported by B^3, Blanc, and CEAF, respectively, whereas a score of 0 in each might reasonably be expected.

The purpose of this paper is to identify contextual features, knowledge resources and lexical patterns for coreference resolution specific to clinical texts, and to evaluate their performance in generating complete coreference chains against the manually annotated gold standard corpora from the 2011 i2b2/VA challenge. We present cross-validation results for each data set within a training set of 589 documents from both corpora, and system performance on a test set of 388 documents. Training data were selected and released to system developers as described in Uzuner et al[11], with test data released 3 months after the training set. Independent evaluation results from the i2b2/VA performance measures[14] are presented for performance of baseline (blind coreference of all mentions of the same class) vs. contextual patterns; both runs are compared using the Mann–Whitney *U* two-sample rank-sum test to determine whether our approach offers significant improvements over baseline. Finally, we compare the results against scores given by a simplified measure that attempts to avoid the anomalous results given by other metrics for null system output.

3. Methods

The basis of our clinical coreference system is a rule-based pipeline that runs within the GATE[21] framework. Rules were developed using the Java Annotation Patterns Engine (JAPE) language, and external domain knowledge integration plugins using Java[1]. JAPE allows pattern matching and evaluation of text *annotations* using a regular expression-like syntax. An annotation represents a marked range in the text, corresponding to some entity or mention, with start and end nodes, a document-unique identifier, and a set of *features* (attributes on the annotation). Each node points to a character offset in the document. One of the benefits of JAPE is that annotations not specified in the input are ignored for pattern matching purposes, which enables patterns to be generalized when, for example, intervening punctuation and prepositions are not significant (this should be clearer in the examples presented).

ODIE corpus mentions had previously been annotated as Pronoun, People, Procedure, DiseaseOrSyndrome, SignOrSymptom, Reagent, LaboratoryOrTestResult, OrganOrTissueFunction, and AnatomicalSite; i2b2/VA corpus mentions as Person, Problem, Treatment, Test and Pronoun. In order to generalize our method across both corpora, and clinical notes in general, these classifications were mapped to three core types: Person, Pronoun and the generic superclass 'Thing' (i.e. clinical terms that are not Person or Pronoun) for the purposes of generalizing the coreference rules (see section 3.5). Our system combines GATE ANNIE[22] text segmentation components with custom named-entity annotators and integration plugins

[1] The domain integration plugins are available from http://vega.soi.city.ac.uk/~abdy181/software/. The complete coreference toolkit is currently being prepared for distribution

developed by the authors to embed clinical domain knowledge and contextual cues into the text, in order to semantically enrich the Person, Pronoun and Thing mentions already present so that potential coreference relations can be computed.

The approach comprises five stages as shown in Fig. 1 and described in detail in sections 3.1 to 3.5 below. In the examples presented, the text delimited by an annotation is shown in square brackets, the annotation type is shown in subscript in initial caps, and annotation features in subscript, lower case. JAPE patterns are shown in an abbreviated form, where token sequences are shown in square brackets, text in curly braces denotes the annotation name, and feature assignment statements are written as `Annotation.feature = value`. For full details of syntax and how to construct JAPE patterns, the reader is referred to Chapter 8 of Cunningham et al.[21]

3.1. Text segmentation

Standard GATE ANNIE[22] components provide initial shallow parsing and phrase chunking. We wrote pattern-matching rules that split the source documents into sections and classify each, based on the text of identifiable headings (such as ‘PHYSICAL EXAMINATION:’ or ‘LABORATORY DATA:’) or paragraph content. Sections, sentences or paragraphs identified as being related to family history or historical lab data were then marked by the system as being potentially excluded from coreference of Thing mentions related to the patient (see section 3.5).

3.2. Identification of supporting entities, context and features

Determining whether two mentions are coreferent or not is usually dependent on the context in which those mentions appear, for example:

`[blood pressure]`$_{\text{Test}}$ `of [100/70]`$_{\text{Measurement}}$ `... [blood pressure]`$_{\text{Test}}$ `of [140/80]`$_{\text{Measurement}}$

`[Patient]`$_{\text{Person}}$ `has a past medical history of [hypertension]`$_{\text{Problem}}$`. [Patient’s mother]`$_{\text{Person}}$ `also has [hypertension]`$_{\text{Problem}}$`.`

The two ‘blood pressure’ Test mentions do not refer to the same external event as they relate to different measurement events, and the two ‘hypertension’ Problem mentions refer to the conditions of two different people.

Identification of entities and mention features that can be used to support or eliminate coreference between two mentions is a key task. Following Zheng et al.[5], selection of features and contextual cues was based on those used by general-purpose coreference systems, and which could be adapted to clinical texts, such as UMLS semantic type agreement, abbreviation expansion, plus additional features identified from a sample of documents from the training corpora.

Table 1 shows the supporting entities and features used for each mention class. To clarify, a supporting entity is a separate annotation identified by the system as providing information relevant to the context in which the Person or Thing mention appears. A supporting feature is something that is either already intrinsic to the mention itself (such as the head word of the noun phrase, or whether the word or phrase is singular or plural), or is the result of storing the text of a nearby supporting entity as a feature on the mention. For example,

`[Mrs Smith]`$_{\text{Person}}$`, a [79-year-old] inpatient of Ward 1`

The text ‘79-year-old’ would be identified as an independent {Age} entity, but that supports the classification of the separate Person mention (see section 3.4), whereas the gender of ‘Mrs

Smith' is a supporting feature, being an intrinsic property of the 'Mrs' honorific, and would be stored as a feature on the 'Mrs Smith' Person mention.

Table 1 Supporting entities and features to identify mention context

Mention class	Supporting entity	Supporting feature
Person	Honorific, FirstName, Surname, GenderIdentifier, Age	role (family, patient, clinician) gender number
Pronoun	VG (verb group), IN (preposition)	gender number case
Thing	Section, Person, Date, Time, Duration, Number, Measurement, Frequency, MedicationRoute, AnatomicalTerm, SpatialConcept, TemporalConcept	number headword laterality (left, right, bilateral) normalizedString (abbreviation expansion, spelling correction, determiner removal) UMLS Concept Unique Identifier (CUI), UMLS preferred name, concept name, semantic type WordNet synonyms, hypernyms, holonyms, meronyms

To identify supporting entities for context, we extended the existing ANNIE Person identifier and wrote a pattern-based recognizer for general named entities such as number, date, time, duration, measurement, name, role, and age using gazetteer lists of primitives and JAPE expressions. In the GATE framework, a gazetteer comprises one or more plain text files (e.g. `anatomy1.lst`) that function as lookup lists, each of which is described in an index file that classifies each list according to major and minor types (e.g. `anatomy1.lst:human_anatomy:location`). The lists themselves comprise one entry per line, where each entry is a term to be looked up in the document, and can be further classified with one or more feature attributes that will be added to the annotation created when a lookup term is found in the document.

For anatomical terms we extracted gazetteer lists of anatomical primitives (parts, spaces, locations, bones, muscles, organs) from Wikipedia[23],[24] and the Foundational Model of Anatomy[25] and wrote JAPE patterns to identify complete anatomical terms in the text via the logical combination of these primitives. To expand and disambiguate abbreviations, we took a list of medical abbreviations from Wikipedia[26], and classified them in a gazetteer according to their corresponding mention classes (e.g. `LPH;term=left posterior hemiblock;type=Problem,DiseaseOrSyndrome`). A JAPE transducer was used to match abbreviations within mentions of the same class (so, for example, ‘PT’ as the content of a Person mention is more likely to mean ‘patient’ rather than ‘prothrombin time’) and store the expanded term as a feature on the mention.

We used MetaMap[27] and the GATE `mmserver` integration plugin[28] to identify term headwords and to add UMLS CUI and UMLS preferred names for each UMLS semantic type identified by MetaMap as features on each Thing mention. To reduce the number of features added, we used MetaMap’s `--term_processing` option (i.e. each mention is treated as a single term), only considered SNOMED CT mappings, and took only the highest-scoring MetaMap mapping group for each mention.

To correct misspellings, we developed a plugin using the GSpell library[29] to provide in-situ correction of misspelt Thing mentions by adding a mention feature containing the suggested spelling. To avoid false positives, spelling correction was limited to words longer than 3 characters, within an edit distance of 1, and only performed on mentions with no MetaMap mapping, and then a MetaMap re-match was attempted on the spell-corrected string.

A normalized string feature, generated from abbreviation expansion, spelling correction and removal of leading determiners and pronouns, was stored as the canonical form for each Thing mention and used for the basis of string comparison (see section 3.5).

To identify general synonyms, hypernyms and holonyms, we developed a plugin that generates WordNet[30] annotations for given input mentions. We used the plugin to pass mention headwords and supporting entities (Table 1) to WordNet, and stored the output as features on the input mention.

The surrounding context of each mention was identified by taking supporting entities within three Tokens either side of the mention, or within the mention itself, and storing this as a feature on the target mention. For example, given this input phrase

> $\texttt{[Culture]}_{\texttt{Test}}$ on blood sample was … $\texttt{[Culture]}_{\texttt{Test}}$ on urine sample was …

we have

> $\texttt{[Culture]}_{\texttt{Test}}$ on $\texttt{[blood]}_{\texttt{AnatomicalTerm}}$ sample
>
> $\texttt{=>Mention}_{\texttt{Test}}$.anatomical_context = blood

[Culture]$_{\text{Test}}$ on [urine]$_{\text{AnatomicalTerm}}$ sample

=> Mention$_{\text{Test}}$.anatomical_context = urine

And

MVA resulted in [3 broken left ribs]$_{\text{Problem}}$ and [1 broken right rib]$_{\text{Problem}}$

gives us

[[3]$_{\text{Number}}$ broken [left]$_{\text{SpatialConcept}}$ [ribs]$_{\text{AnatomicalTerm}}$]$_{\text{Problem}}$

=>Mention$_{\text{Problem}}$.spatial_context = left,

Mention$_{\text{Problem}}$.anatomical_context = rib

[[1]$_{\text{Number}}$ broken [right]$_{\text{SpatialConcept}}$ [rib]$_{\text{AnatomicalTerm}}$] $_{\text{Problem}}$

=>Mention$_{\text{Problem}}$.spatial_context = right,

Mention$_{\text{Problem}}$.anatomical_context = rib

3.3. Pronoun classification

Using string matching and surrounding part-of-speech (POS) tags, we developed a general-purpose classifier in JAPE to categorize pronouns according to type (anaphoric or pleonastic); case: nominative (*I, he, she*); objective (*me, him*); possessive (*my*); reflexive (*myself*); nominative-possessive (*mine, hers*); number (singular or plural); class: Person (personal pronouns), Thing (*it*, *that*, *these*, *those* etc), Location (*here, there, where*); person (first, second, third); and gender. Third-person plural pronouns (*they, their, them*) were not categorized at this stage as their assignment (Person or Thing) is context-dependent.

Only anaphoric pronouns will participate in coreference, so pleonastic ‘it’ and ‘that’ references are identified using a set of general patterns that look for temporal phrases, verb ‘to be’ phrases ending in ‘that’ or ‘whether’ (e.g. ‘*It is unclear whether …*’, ‘*it is important to note that …*’) and modal ‘to be’ phrases ending in an infinitive or a preposition (e.g. ‘*It should be possible for …*’, ‘*It may be sensible to consider …*’). JAPE expressions for these patterns, with accompanying examples for clarity, are shown below (where `|` `?` and `(n, m)` denote regular expression occurrence operators):

```
[“It”]    {VG contains [“be”]}({Day}|{Date}|{Time})

It        is                         Tuesday

It        was                        10pm


[“It”]    {VG.type == modal?, VG contains [“be”]}

It        is

It        may be

It        should be


({ADV}(0,2) {ADJ})(0,3)

somewhat unclear

important

possible


( {VG.tense == infinitive}?   [“whether|if|that”] )  |  {IN}
```

```
                                     whether

    to note                          that

                                                                 for
```

where VG= verb group, ADV= adverb, ADJ=adjective, IN=preposition.

3.4. Person and personal pronoun categorization

Coreference systems for general English texts typically make use of gender, number and grammatical role information to resolve coreference of personal pronouns. A pattern expressing possible pronominal coreference between a person and personal pronoun within the same sentence or between consecutive sentences might then be written as:

$\{\texttt{Mention}\}_{\texttt{Person}, \textit{gender, number, grammar\_role}}\ (!\{\texttt{Mention}\}_{\texttt{Person}})+$
$\{\texttt{Mention}\}_{\texttt{Pronoun}, \textit{gender, number, grammar\_role}}$

i.e. ‘match a Person mention followed by a Pronoun mention where there are no intervening Person mentions’, and where Person.gender=Pronoun.gender, Person.number=Pronoun.number, and Person.grammar_role=Pronoun.grammar_role.

For example

$[[\texttt{Jane}]_{\texttt{FirstName,female}}\ [\texttt{Smith}]_{\texttt{Surname}}]_{\texttt{Person,female,singular,subject}}$ has a past history of [hypertension]. $[\texttt{She}]_{\texttt{Person,female,singular,subject}}$ was admitted on ...

A typical system might also match occurrences of congruent name strings such as ‘Smith’, ‘Jane’, ‘Ms Smith’.

However, in anonymized clinical notes, the deidentification process potentially loses any link between the person's name and their gender, or between initial and subsequent mentions. Does 'XXXX' annotated as a Person refer to the patient, and are they male or female? Does Mr XYXY refer to the same person? Additional classification steps need to be employed to discriminate these cases.

For example, phrases extracted from the training corpus that identify the gender of the patient tend to be of the form:

[Patient]$_{Person}$ is a 40-year-old male with [type 2 diabetes]$_{Problem}$

[XXX]$_{Person}$ is an 80 y/o female admitted on …

[This]$_{Person}$ is a baby boy born on …

Which can be generalized to the pattern:

{Mention} { VG contains ["be"]} {Age} {GenderIdentifier} ({VG} | {Mention}$_{Problem}$)

=> Mention.class=Person, Mention.semantic_role = patient, Mention.gender = GenderIdentifier

In general, it seems reasonable to assume that the key protagonist in a clinical note is the patient, and that the actions described in the note will center around them. Analysis of the manually annotated coreference chains in the 589 training documents confirmed this: 86% of all personal pronoun mentions (*he, she, his, her* etc) referred to the patient, and 75% of all Person mentions also referred to the patient. The remaining mentions referred to members of the clinical team, to family/significant others, or to the person receiving the report. Therefore we classified Person and personal Pronoun mentions according to 3 main types:

patient

patient's family or significant other

clinician

- author
- attending
- receiver
- referred clinicians (e.g. external teams, social workers etc)

Classification was performed using lexical rules and gazetteers of family relations (*wife, daughter, brother* etc), clinical roles and honorifics (*physician, doctor, nurse*, *Dr.*, *M.D.*, etc) and contextual cues (e.g. section heading content and gender identifiers). Nominal Person mentions were classified as referring to the patient by default, unless the context suggested one of the other categories. For example, verb roots associated with a clinician include '*consult*', '*attend*', '*dictate*', and certain past participles relate different protagonists, i.e.

```
{Mention}Person, semantic_role1 ["seen|treated|evaluated|treated…"]VG
["by"] {Mention}Person, semantic_role2
=> Mention.semantic_role1 = patient, Mention.semantic_role2
= clinician
```

and

```
{Mention}Person, semantic_role ["performed|signed|verified…"]VG
```

```
    => Mention.semantic_role = 'clinician'.
```

and more generally, using role identifiers:

$$\texttt{\{Mention\}}_{\texttt{Person, }\textit{semantic\_role}}\ \texttt{\{RoleIdentifier\}}_{\textit{type}}$$

```
    => Mention.semantic_role = RoleIdentifier.type
```

Personal pronouns were considered as having either global or local scope. By default, personal pronouns outside quoted speech have global scope. Second- (*you, your*) and third-person (*he, she*) singular pronouns are provisionally assigned to the patient if the pronoun's gender matches that of the patient. In the absence of gender cues, the document frequency of male and female pronouns were used to infer the patient's gender, given the prior probability (86%) that a personal pronoun refers to the patient. First-person pronouns (*I*, *we* etc) are assigned to the report's author.

Local scope exceptions are then identified as follows:

A context switch triggered by a possessive pronoun, e.g. '*his wife ... she*', '*his oncologist ... he*'. Additionally, the locally scoped pronoun should agree in gender with that of the new context, if present.

A context switch triggered by the appearance of a new actor, e.g. '*the social worker is Barbara Cole. She can be contacted on ...*' Again, gender features should agree, if present.

Role of the report's receiver: By default, references to *you*, *your* etc are assumed to be directed to the patient, unless it is clear that the recipient is a clinician (e.g. '*your patient*'), in which case, the second-person pronoun is assigned a clinical role.

3.5. Coreference resolution

Coreference resolution rules follow similar heuristics to the multi-pass sieve recently presented by Lee et al.[7] for newswire text, but with specific consideration of world and clinical domain knowledge. While Lee et al. resolve pronouns on a final pass, we resolve pronominal coreference for each mention class first, and each potential mention-pair is considered only once, as described below. Furthermore, we address some of the weaknesses of the traditional mention-pair approach, by making use of the contextual information surrounding each mention and/or pronoun, and by making use of centering theory to give preference to coreferents that grammatically agree with forward-looking centers[4]. Briefly, centering theory suggests that, in a coherent discourse, entities and their coreferent pronouns will occupy the same grammatical position in the sentence or clause – usually that of the subject where there is a single entity, but also in parallel subject/object pairs in the case of two or more entities and pronouns, as in:

> [Patient]$_{\text{subject}}$ suffers from [lower back pain]$_{\text{object}}$. [He]$_{\text{subject}}$ takes [Vicoprofen]$_{\text{indirect\_object}}$ for [this]$_{\text{object}}$.

Additionally, *protagonist theory*[31] suggests that narrative events are centered on one or more key actors. Coreferring actors share congruent verbs, and distinct sets of verbs are typically associated with different actor types. Narrative events can therefore be identified by a common protagonist and associated verbs[31]. By inference, a set of narrative events (e.g. the admission, assessment, test and treatment process documented in clinical notes), verbs and protagonists, should facilitate identification of coreferent mentions.

Taking the set of all mentions, we create subsets according to mention class, and within each subset, compare pairs of mentions in document order. For example, the first Treatment mention will need to be tested against all following Treatment mentions, the second against the third, fourth etc. For a given subset, the maximum number of comparisons that need to be made, for each mention class, is given by

$$\sum_{i=1}^{n-1} (i - 1) \qquad = 1/2n(n-1) \qquad \approx \mathrm{O}(n^2)$$

where n is the number of mentions in the class.

However, for efficiency, each input subset is pruned of successful mention pairings during traversal, which should reduce the computational overhead of comparing large numbers of mentions. This is illustrated in Figure 2, which represents a document containing two classes of mention, the selection of one class of mention and the coreference iteration process. As shown in the figure, when the mention pointed to by the outer iterator matches a mention pointed to by the inner iterator, the features of the former are cloned to the latter, the outer iterator points to the coreferent mention, and the inner iterator is incremented to the next mention. Once the inner iterator completes, coreferent mentions are pruned and the process repeats until the outer iterator completes.

A set of linked lists corresponding to each coreference chain is thus created, where each mention is assigned a unique identifier and, for each link in the chain, we store the annotation id of the coreferent on the antecedent (and a back reference from the coreferent to the antecedent is created, to form a double-linked list). This allows *in situ* evaluation via the GATE corpus quality assurance toolkit[21], by testing the value (or null, for singletons) of the coreference identifier on each mention, which should agree between the key set and system output. Each linked list can then be traversed and serialized to a coreference chain text file, for evaluation via external metrics, such as those used by the i2b2/VA evaluation script[14].

3.5.1. Person coreference chain generation

Following the addition of the above-described classification features to Person and Pronoun mentions, pairs of these mentions are traversed in document order and compared according to the following rules:

1. Strings are normalized by removing leading determiners and pronouns
2. 'Who' pronouns are paired with the immediately preceding Person mention.
3. Pairs of nominal Person–third person-pronominal mentions are coreferenced if their genders (if present), scope, role/type and number (singular or plural) agree. Uncategorized third-person plural pronouns were coreferenced with plural Person mentions (e.g. *the paramedics*) with grammatical role agreement in the absence of interventing plural Thing mentions. The features of the antecedent are cloned to the coreferent pronoun, so that the pronoun is now effectively a nominal Person mention, and the matching process continues from nominal to pronominal.
4. Person mentions classified as 'patient' are coreferenced if the genders agree. Person mention pairs classified as 'family' are coreferenced if the genders agree and the string values or WordNet synonyms agree (e.g. *sister* will corefer with *sibling*). Other Person mention pairs are coreferenced by evaluating the following, in order:
    a. Exactly matching name strings are coreferenced
    b. Mentions with matching first names and surnames, where identifiable, are coreferenced
    c. First-person pronouns of global scope are coreferenced and linked to the primary clinician (usually the report's author)

d. Approximately matching strings over 4 characters long are coreferenced. Using the SecondString Java library[32], and following Cohen et al.[33] we take the mean value of the Jaro-Winkler[34] and Monge-Elkan[35] string comparison metrics, which returns a value between 0 (no match) and 1 (strong match). If the result exceeds a tunable threshold (we use 0.85[†]), the two strings are coreferenced. This step allows de-identified name pairs such as '***NAME[AAA , BBB]*' : '***NAME[AAA]*', and '*Mr. BBBBB*' : '*BBBB*' to be coreferenced.

The following example demonstrates this process:

> [Mr WWWWW] is a 58 y/o gentleman [who] was admitted … by [Dr FFFF]. … [He] was assessed by [Dr GGGGG] … [She] has referred [WWW] to [the orthopedics team]; [he] will be followed up by [them].

Following the feature identification and classification described in section 3.4 above, we have

> [Mr WWWWW]$_{\text{Person,patient,male,singular}}$ is a 58 y/o gentleman [who]$_{\text{Person}}$ was admitted … by [Dr FFFF]$_{\text{Person,clinician,singular}}$. … [He]$_{\text{Person,patient,male,singular}}$ was assessed by [Dr GGGGG] $_{\text{Person,clinician,singular}}$ … [She]$_{\text{Person,female,singular}}$ has referred [WWW]$_{\text{Person,patient,male,singular}}$ to [the orthopedics team]$_{\text{Person,clinican,plural}}$; [he]$_{\text{Person,patient,male,singular}}$ will be followed up by [them]$_{\text{Person,plural}}$.

After steps 1-4 above, we have

[†] The value of 0.85 was determined by examining the Jaro-Winkler and Monge-Elkan scores on a selection of 65 randomly selected coreferent and non-coreferent mention pairs from the training set; lower values tended to accept false positives, higher values false negatives.

[*Mr WWWWW*]$_{\text{Person,patient,male,singular}}$ is a 58 y/o gentleman [*who*]$_{\text{Person,patient,male,singular}}$ was admitted … by [Dr FFFF]$_{\text{Person,clinician,singular}}$. … [*He*]$_{\text{Person,patient,male,singular}}$ was assessed by [Dr GGGGG]$_{\text{Person,clinician,singular}}$ … [She]$_{\text{Person,clinician,female,singular}}$ has referred [*WWW*]$_{\text{Person,patient,male,singular}}$ to [**the orthopedics team**]$_{\text{Person,clinican,plural}}$; [*he*]$_{\text{Person,patient,male,singular}}$ will be followed up by [**them**]$_{\text{Person,clinician,plural}}$.

where Person coreference chains are indicated via corresponding levels of emphasis.

3.5.2. ‘Thing’ coreference chain generation

Coreference of general clinical terms follows a similar approach as for Person mentions. Anaphoric pronouns of class Thing (see section 3.3) are resolved against the most recent Thing antecedent with the same grammatical role (e.g. subject, object, indirect object), followed by the cloning of antecedent features to the anaphor so that the anaphor is converted to a nominal mention. Uncategorized third-person plural pronouns were coreferenced with plural Thing mentions (e.g. *the sutures*) with grammatical role agreement in the absence of intervening plural Person mentions.

Nominal coreference is then attempted for pairs of mentions of the same class, in document order. This is more complex than for Person mentions and involves a voting process based on the number of matching features identified from rules given in the i2b2/VA coreference annotation guidelines,[13] and the ODIE anaphoricity annotation guidelines[36]. In summary, these rules are:

For Thing mentions of the same class, consider pairing if:

1. mention synonyms refer to the same episode. For example, '*chills*' with '*shivering*' and '*inflammation*' with '*swelling*', if other contexts are equal;

2. a mentions occurs with its hypernym and if both refer to the same episode. For example, '*staph bacteraemia*' with '*the [infection]*$_{hypernym}$', '*stereotactic biopsy*' with '*the [procedure]*$_{hypernym}$', '*dyspnea*' with '*shortness of breath*' (UMLS preferred name), '*CABG*' with '*the revascularization*';

3. there is a holonym/meronym relation between anatomical terms within or surrounding mentions;

4. there is agreement between the headwords of mention noun phrases where the antecedent is more specific than the coreferent, where all other contexts are equal. For example, '*intermittent right neck [swelling]*$_{headword}$' with '*the [swelling]*$_{headword}$'.

Consider eliminating pairing where:

5. spatial concepts within each mention are different. For example, '*chronic [bilateral]*$_{SpatialConcept}$ *lower extremity swelling*' should not be coreferenced with '*the [right]*$_{SpatialConcept}$ *lower extremity swelling*';

6. the quantitative, temporal or anatomical context around each mention are different. For example: '*[2017-06-14 02:06AM]*$_{TemporalConcept}$*: WBC - 9.4*' vs. *'[2017-06-13 08:05PM]*$_{TemporalConcept}$*: WBC - 9.4*' and '*blood pressure of [120/80]*$_{Measurement}$*'* vs. '*blood pressure' of [100/70]*$_{Measurement}$'. Also '*simple*

atheroma in the [aortic root]$_{AnatomicalTerm}$' vs. '*simple atheroma in the [ascending aorta]*$_{AnatomicalTerm}$'.

7. Either mention is within a sentence or section of the document related to family history.

Coreferencing is not attempted if either of the mention pair occurs in an excluded section (rule 7) or if the contexts do not match (rules 5 and 6). A context match between mentions is made if there is a direct match between contextual features on both mentions (see section 3.2) or there is a whole/part relation between the anatomical contexts of both mentions.

If contexts match, or the antecedent mention has a contextual feature and the potential coreferent does not, then

a. If there is an exact match between normalized strings (see section 3.2), the coreference is marked and iteration continues with the next mention pair.

b. Otherwise, consider marking a match if one or more of the following are true, in order of preference:

    i. The UMLS CUIs of the head word/phrase in each mention match, or if there is intersection between sets of headword CUIs (where there is more than one), and the spatial contexts (e.g. left, right);

    ii. There is intersection between sets of anatomical terms within each mention and between sets of UMLS semantic types for the headword/phrase;

    iii. The headwords and anatomical contexts match;

    iv. There is an approximate string match, as measured by the mean Jaro-Winkler/Monge-Elkan score within the defined threshold (see 3.5.1).

4. Results

Summary validation results across all mention classes for the training portion of the i2b2/VA (492 documents) and ODIE corpora (97 documents) are reported in Table 2. Detailed results against the withheld test portions of the i2b2/VA (322 documents) and ODIE corpora (66 documents) are shown in Table 3. For each mention class, micro-average recall, precision and *F*-measure scores across the B^3, MUC and CEAF scores output by the i2b2 coreference evaluation script, and the pairwise mention/feature matching metric of the GATE corpus QA toolkit, are shown. Results for the baseline, which involves blind coreference of all mentions of the same class (Test, Treatment, DiseaseOrSyndrome etc) into a single chain, are shown in italics (test set only). Mann–Whitney *U* two-sample rank-sum test results for matched pairs of baseline–system performance results are shown at the end of each column of Table 3.

In the gold-standard test data, the number of mentions, chains, mean and maximum coreference chain lengths were 3002, 419, 5.7 and 90 for the ODIE corpus; and 43867, 5277, 4.3 and 122 for the i2b2/VA corpus. The mean number of true mentions, chains and coreference relations per document were 45.5, 6.4 and 36.2 for the ODIE corpus, and 136.3, 16.4 and 70.5 for the i2b2/VA corpus. In the absence of analysis of the class distribution of mentions in the true chains, the likelihood that a given mention will appear in a coreference chain can be estimated as

$$p_c = N_{\mathrm{chains}} * \mu_c / N_{\mathrm{mentions}}$$

where N_{chains} is the total number of true chains in the gold standard, μ_c is the mean chain length and N_{mentions} the total number of true mentions. For the ODIE corpus, this gives $p_c \approx 0.8$, for the i2b2/VA corpus $p_c \approx 0.5$.

Coreference chain length varied widely between document type: discharge and progress reports from both corpora had higher mean (5.42) and maximum chain length (106) than radiology, surgery and pathology reports (mean 3.61, maximum 18).

As shown in Table 3, the coreference-specific metrics show a wider discrepancy between baseline and system performance than the GATE pairwise evaluation metric, particularly in relation to system recall. Also, the GATE metric reported reduced system performance over the baseline for five classes (Disease, AnatomicalSite, OrganOrTissueFunction, Procedure, LaboratoryOrTestResult), in contrast to the other metrics, which reported improved system performance *F*-measure over the baseline for all classes apart from OrganOrTissueFunction.

Table 2. Training corpus coreference evaluation - summary results

Corpus	**Micro-average over i2b2/VA metrics***			**Micro-average over GATE QA metrics**		
	Precision	*Recall*	*F*	*Precision*	*Recall*	*F*
I2b2/VA	**0.905**	**0.855**	**0.878**	**0.923**	**0.923**	**0.923**
ODIE	**0.771**	**0.828**	**0.796**	**0.765**	**0.765**	**0.765**
Reagent†	*0.352*	*0.160*	*0.131*	0.00	0.00	0.00

* Unweighted average of MUC, B^3 and CEAF scores according to i2b2/VA evaluation script[14].

† Zero system results for this class: shown to highlight anomalous scores reported by existing metrics in comparison to GATE QA metric (see Discussion).

Table 3. Test corpus coreference evaluation results, baseline (italics) vs system

	Micro-average over i2b2/VA metrics*			**Micro-average over GATE QA metrics****		
i2b2/VA corpus	*Precision*	*Recall*	*F*	*Precision*	*Recall*	*F*
All classes	***0.771***	***0.492***	***0.519***	***0.810***	***0.810***	***0.810***
	0.895	**0.857**	**0.875**	**0.915**	**0.915**	**0.915**
Person	*0.739*	*0.641*	*0.593*	*0.800*	*0.800*	*0.800*
	0.870	0.866	0.868	0.900	0.903	0.903
Test	*0.404*	*0.354*	*0.166*	*0.930*	*0.930*	*0.930*
	0.849	0.756	0.792	0.920	0.960	0.943
Treatment	*0.520*	*0.425*	*0.306*	*0.770*	*0.770*	*0.770*
	0.848	0.801	0.822	0.880	0.903	0.893
Problem	*0.459*	*0.414*	*0.249*	*0.770*	*0.770*	*0.770*
	0.860	0.790	0.821	0.888	0.918	0.900
Mann-Whitney two-tailed, $n_1 = n_2 = 5$	$U = 25$ $p = 0.012$	$U = 25$ $p = 0.012$	$U = 25$ $p = 0.012$	$U = 20$ $p = 0.144$	$U = 21$ $p = 0.095$	$U = 21$ $p = 0.095$
ODIE corpus						
All classes	***0.729***	***0.624***	***0.653***	***0.770***	***0.770***	***0.770***
	0.765	**0.827**	**0.792**	**0.780**	**0.780**	**0.780**
People	*0.758*	*0.701*	*0.719*	*0.780*	*0.780*	*0.780*
	0.756	0.797	0.769	0.820	0.820	0.820
Disease	*0.649*	*0.533*	*0.555*	*0.770*	*0.770*	*0.770*
	0.672	0.758	0.709	0.730	0.760	0.750
Symptom	*0.677*	*0.478*	*0.476*	*0.770*	*0.770*	*0.770*
	0.837	0.812	0.824	0.860	0.890	0.880
Anat. Site	*0.731*	*0.591*	*0.616*	*0.790*	*0.790*	*0.790*
	0.671	0.745	0.703	0.650	0.660	0.650
Reagent[†]	-	-	-	-	-	-
	-	-	-	-	-	-
Organ Fn	*0.606*	*0.602*	*0.596*	*1.00*	*1.00*	*1.00*
	0.426	0.542	0.474	0.710	0.830	0.770
Lab Result	*0.386*	*0.425*	*0.353*	*0.950*	*0.950*	*0.950*
	0.610	0.579	0.590	0.900	0.950	0.920
Procedure	*0.690*	*0.555*	*0.582*	*0.800*	*0.800*	*0.800*
	0.714	0.804	0.751	0.710	0.760	0.730
Mann-Whitney two-tailed, $n_1 = n_2 = 9$	$U = 35.5$ $p = 0.690$	$U = 63.5$ $p = 0.047$	$U = 59.5$ $p = 0.103$	$U = 30.0$ $p = 0.379$	$U = 36.5$ $p = 0.757$	$U = 31.5$ $p = 0.453$

Scores in italics represent the baseline, which consists of coreferencing pairs of mentions of the same class (Test, Treatment etc) into a single chain.

* Unweighted average of MUC, B^3 and CEAF scores according to i2b2/VA evaluation script[14].

** Results for ‘All classes’ account for singleton pronouns and thus differ from the mean over all classes shown.

† No mentions in key or system set.

5. Discussion

The recent review by Zheng et al.[5] called on research into the portability of general coreference resolution methods to the clinical domain. We have combined these methods with additional patterns to address weaknesses in the general approaches when applied to clinical notes, namely integration of external domain knowledge, dealing with name deidentification/anonymisation, spelling errors and inconsistencies, use of abbreviations, and wide scope of pronominal resolution.

As measured by the coreference-specific metrics, system performance on the i2b2/VA corpus shows significant improvement in precision, recall and *F*-measure over the baseline (p<0.05). However, for the ODIE corpus, only recall is significantly improved (p<0.05), while precision and *F*-measure are improved, but not significantly (p>0.05). While initially surprising, the results may be explained as follows. In the gold-standard ODIE data, around 80% of all mentions are in a coreference chain, but a typical ODIE document contains only about 6 such chains. So a baseline coreference of simply chaining, in each document, all mentions of the same class, has a reasonable chance of success (*F*=65.3%). However, for the i2b2/VA gold standard, only 50% of mentions are in a coreference chain, yet there are on average about 16 chains per document, so the baseline method should perform less well, which it does (*F*=51.9%).

For the system output, individual errors will have a greater impact on overall accuracy in documents with fewer anaphoric relations than in those with many relations. This was typically the case with the ODIE corpus (on average 36.2 relations per document vs. 70.5 for the i2b2/VA corpus), which also had a higher mean chain length (5.7 vs. 4.3). These may partially explain the overall weaker ODIE results in comparison to those for the i2b2/VA corpus, although further work is needed to analyse performance in relation to coreference chain length. In addition, it has been suggested that some coreference evaluation metrics, favour longer coreference chains[5].

Overall, however, the results suggest that the presented approach, which augments generic methods (based on headword and pronoun-matching rules using gender, role, number and recency agreement) with external domain knowledge resources, plus consideration of quantitative, spatial, temporal, and anatomical modifiers, provides greatly increased coreference resolution performance over general-purpose tools (where *F* ranges from 0–35%)[8]. In evaluating the performance of these tools[6][7], Hinote et al.[8] used the same corpora and coreference-specific evaluation metrics as our system.

With some qualifications, our method also appears to offer an improvement over a number of previously reported clinical coreference systems[9][10][19]. Romauch[9] used a corpus of clinical guideline documents and did not detail the evaluation metrics used, so results may not be directly comparable. He[10] used a small corpus of 47 discharge summaries that may be similar to those in the i2b2/VA corpus, and reported scores from the B^3 and MUC metrics used in the current study, so comparison with the current results seems reasonable. Zheng et al.[19] reported results on a subset of the ODIE corpus used here and used the same evaluation metrics. However, their system performed end-to-end identification and coreference of clinical terms, whereas our system (as with [8]-[10]) performs coreference only on existing mentions. Zheng et al. estimated that errors in term recognition accounted for ~20% of system errors; it may be that extending our system to provide end-to-end evaluation would lead to a similar reduction in performance.

System results were submitted to the 2011 i2b2/VA Natural Language Processing Challenge for Clinical Records[11], where it ranked overall 7th out of 28 submissions to the ‘coreference only’ tracks. Precision against the i2b2/VA corpus was equal to that of the top-performing systems; for a full comparison, see Uzuner et al[11]. More importantly, perhaps, our system appears to perform at least as well as human annotators – results for the ODIE corpus are comparable to the mean inter-annotator agreement (IAA) reported[12] of 75.4%, compared to our system performance of 79.2% (IAA scores were not available for the i2b2/VA corpus).

Performance on both the training and test data was in close agreement (i2b2/VA metrics: 79.6% *vs.* 79.2% for ODIE; 87.8% *vs.* 87.5% for i2b2/VA; GATE QA tool metric: 76.5% *vs.* 78.0% for ODIE, 92.3% *vs.* 91.5% for i2b2/VA), which suggests that the rules for feature extraction and coreference resolution were not over-fitted to the training set. However, error analysis revealed three areas where system performance might be improved:

1. *Errors of commission or omission*: For Person mentions, these resulted from incorrect categorization by the system. For other classes, errors occurred where contextual cues had been incorrectly identified, or where the string similarity metrics had reported a false match or lack of match. Spurious pronominal coreferences occurred where pleonastic it/that pronouns had been incorrectly classified as anaphoric.

2. *Broken coreference chains*: Coreferences were correct, but were reported across 2 or more chains, when a single chain should have been reported.

3. *Determinist behaviour*: Unlike machine learning approaches, deterministic rules cannot model ground truth inconsistencies. In 28 of the 46 Beth Israel records in which the attending physician was annotated, the 'Attending' heading and physician name following were coreferenced. In the remaining 18, they were not. There were other inconsistencies in the coreferencing of names with their clinical role in both corpora. However, our deterministic

rules did not allow for such inconsistencies, and always coreferenced physician names with their clinical role.

The system performed well at coreferencing Person mentions across all document types. For the ODIE corpus, the system was weak at coreferencing AnatomicalSite and OrganOrTissueFunction mentions, and the baseline performed better. Pathology reports in particular were problematic, requiring more domain knowledge than we had embedded in the system. For example, the ability to coreference carcinoma mentions that are linked to the formation of a mass, or pairing histological studies such as ‘chemical stains’ with ‘MLH1’. Similar domain knowledge resource limitations were also noted by He[10]. Further work would be required to determine in detail the contribution made by each of the domain knowledge resources (Gspell, WordNet, MetaMap, Wikipedia, Foundational Model of Anatomy) to the performance of the system.

Our system does not impose a limit on the distance between coreferents. In contrast, Zheng et al.[19] imposed a 10 sentence window, as a sample of the training data suggested that a larger limit led to an unacceptable reduction in precision. However, they found that this limit was the most frequent source of recall error, as coreference relations can often span large distances, for example, between the History of Present Illness and Final Diagnosis sections at opposite ends of the document. Therefore, further work could involve examining the effect of varying the distance limit between mentions on the precision and recall of our system.

Our acyclic, forward linked-list approach to coreference chain generation ensures that a given mention only participates in a single coreference chain. By cloning antecedent features to the anaphor and the use of a double-linked list, all coreference relationships can be extracted from any starting node, which simplifies identification of transitive closure and reduces the complexity of the task.

The approach allows performance evaluation via simple comparison, between the key set and system output, of individual nodes and their link identifiers. That is, for each mention that exists in both key set and system output, start/end offsets and coreference id must match. This pairwise evaluation is similar to the MUC metric[20], but unlike MUC it does, however, lead to strict scoring of coreference chains; for example where the key set is A→B→C (transitively, A is coreferent with C), a system output of A→C would be penalised for the missing B link, despite correctly marking the transitive closure.

As with the B^3 metric, but unlike the MUC metric[20], this method also takes into account singletons (i.e. those with null coreference id). Unlike B^3, twinless mentions (those that exist in one set but not in another) are dealt with in the same way as for evaluation of named entity recognition – i.e. mentions in the system set, but not in the key, penalise precision, and mentions in the key, but not in the system, penalise recall. This is an important point – the GATE QA metric is a general purpose tool that we have configured for coreference evaluation: it gives equal weight to correctly identified twinless mentions as it does to coreference pairs, and conversely, equal penalty to incorrectly identified twinless mentions and coreference pairs. This may explain the overall higher scores reported for the GATE QA metric in comparison to the coreference-specific metrics, and its apparent lower sensitivity and specificity in scoring the baseline vs. system results, where it records no significant performance improvement over the baseline, in contrast to the other metrics. The GATE QA metric does, however, return an *F*-measure of 0 for null system output, rather than the positive results given by other metrics (for example, the ‘Reagent’ class in Table 2).

For both corpora, there was a notable divergence in the scoring of OrganOrTissueFunction, Test and LaboratoryOrTestResult performance as measured by both evaluation methods (i2b2/VA metrics: low or lowest scoring; GATE QA metric: highest scoring in system or baseline). These classes, however, appeared with the lowest frequency of all classes, and rarely participated in coreference in the gold standard, suggesting that sparsely populated coreference chains are not handled well by existing metrics.

5.1. Limitations

At present, the system performs coreference on existing clinical mentions; it requires that at least 'Thing'-type concepts (such as Procedure and DiseaseOrSyndrome) have been annotated by a previous step. However, it should be straightforward to extend the system to provide end-to-end annotation and coreference by adding a MetaMap[27] preprocessing step to the pipeline that provides this initial classification of noun phrases and prepositional phrases according to their UMLS semantic type, or using one of the recently described clinical concept recognition systems (e.g. [37]).

Simple string matching and POS-based pronoun classification cannot easily disambiguate third-person plural pronouns (section 3.3). Although this turned out not to be a major problem in the current corpora (post-hoc analysis of the training data showed that these comprised ~1% of all personal pronoun mentions, and 79% of these were coreferenced with a Person in the gold standard), this is a potential source of error, and requires modification of the pronominal coreference rules.

Hand-crafted, procedural rules to classify Person and Pronoun mentions and to process the features extracted by the system to generate coreference chains may be be hard to maintain.

Instead, these features could be used as the input to a supervised learning process (such as a mention-pair classifier or cluster ranking model) to augment or replace these rules.

Although the results suggest that the patterns demonstrated here are reasonably generalizable across the 977 documents that came from a wide variety of sources, counter-examples can doubtless be found. Further work should investigate performance on clinical notes from other centers to determine the generalizability of our approach.

6. Conclusion

Generation of coreference chains provides a means to extract linked narrative events from clinical notes. A novel method of generating coreference chains using progressively pruned linked lists has been demonstrated that reduces the search space and facilitates evaluation by a number of metrics. The GATE QA metric can be configured to provide a rough-and-ready evaluation of coreference performance during system development, but lacks the sensitivity of coreference-specific metrics, although it avoids some of their anomalous results in certain circumstances.

Several patterns, features and knowledge integration components for resolving coreference resolution in clinical notes have also been presented. These components have been developed as independent modules and integrated into an information extraction pipeline. System output has been independently evaluated, and performance exceeds that of general purpose tools, and is comparable to that of recently reported, state-of-the art systems.

Some components are now available to the research community from http://vega.soi.city.ac.uk/~abdy181/software/ and the complete pipeline is currently being prepared for distribution so that it can be evaluated by others on new data sets.

7. Acknowledgements

Deidentified clinical records used in this research were provided by the i2b2 National Center for Biomedical Computing funded in part by the grant number 2U54LM008748 from National Library of Medicine, and were originally prepared for the 2011 i2b2/VA Challenges in Natural Language Processing for Clinical Data Shared Task supported by 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. Phil Gooch acknowledges funding and support from the Engineering and Physical Sciences Research Council (EPSRC) in carrying out this research as part of PhD studentship EP/P504872/1.

8. References

[1] Gasperin C. Statistical anaphora resolution in biomedical texts. University of Cambridge Computer Laboratory Technical Report No. 764. 2009. Available from: http://www.cl.cam.ac.uk/techreports/UCAM-CL-TR-764.pdf [Accessed 31 January 2012]

[2] van Deemter K, Kibble R. On coreferring: Coreference in MUC and related annotation schemes. Computational Linguistics. 2001;26(4):629-637.

[3] Rahman A, Ng V. Narrowing the modeling gap: a cluster-ranking approach to coreference resolution. Journal of Artificial Intelligence Research. 2011;40:469–521.

[4] Ng V. Supervised noun phrase coreference research: the first fifteen years. Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics 2010, pp. 1396–1411. Stroudsburg, PA: Association for Computational Linguistics.

[5] Zheng J, Chapman WW, Crowley RS, Savova GK. Coreference resolution: A review of general methodologies and applications in the clinical domain. J Biomed Inform. 2011; 44(6):1113-22.

[6] Versley Y, Ponzetto SP, Poesio M, Eidelman V, Jern A, Smith J, Yang X, Moschitti A. BART: A Modular Toolkit for Coreference Resolution. Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC 2008).

[7] Lee H, Peirsman Y, Chang A, Chambers N, Surdeanu M, Jurafsky D. Stanford's multi-pass sieve coreference resolution system at the CoNLL-2011 Shared Task. Proceedings of the CoNLL-2011 Shared Task, 2011.

[8] Hinote D, Ramirez D, Ping Chen. A comparative study of co-reference resolution in clinical text, The Fifth i2b2/VA/Cincinnati Workshop on Challenges in Natural Language Processing for Clinical Data, Washington DC, October, 2011.

[9] Romauch M. Coreference resolution in clinical practice guidelines. Diplomarbeitspräsentationen der Fakultät für Informatik, 2009. Wien: Technische Universität Wien.

[10] He T-Y. Coreference resolution on entities and events for hospital discharge summaries. Thesis (M. Eng.), Massachusetts Institute of Technology, Dept. of Electrical Engineering and Computer Science, 2007.

[11] Uzuner Ö, Bodnari A, Shen S, Forbush T, Pestian J, South B. Evaluating the state of the art in coreference resolution for electronic medical records. J Am Med Inform Assoc. Published online first: 24 February 2012 doi:10.1136/amiajnl-2011-000784.

[12] Savova GK, Chapman WW, Zheng J, Crowley RS. Anaphoric relations in the clinical narrative: corpus creation. J Am Med Inform Assoc. 2011;18(4):459-65.

[13] Uzuner O. 2011 i2b2/VA coreference annotation guidelines for the clinical domain. Available from: https://www.i2b2.org/NLP/Coreference/assets/CoreferenceGuidelines.pdf [Accessed 31 January 2012].

[14] Bodnari A. Coreference resolution evaluation script, 1.6.3, June 2011. Available from: https://www.i2b2.org/NLP/Coreference/assets/coreference_evaluation_metrics.zip [Accessed 29 January 2012].

[15] Bagga A, Baldwin B (1998). Algorithms for scoring coreference chains. Proceedings of the 1st International Conference on Language Resources and Evaluation, Granada, Spain, 1998;563–566.

[16] Vilain M, Burger J, Aberdeen J, Connolly D, Hirschman L. A model-theoretic coreference scoring scheme. Proceedings of the 6th Message Understanding Conference (MUC-6). 1995;45–52. San Mateo, CA: Morgan Kaufmann.

[17] Luo X. On coreference resolution performance metrics. Proceedings of the Human Language Technology Conference and the 2005 Conference on Empirical Methods in Natural Language Processing, Vancouver, B.C. 6–8 October 2005, pp. 25–32.

[18] Recasens M, Hovy E. BLANC: Implementing the Rand index for coreference evaluation. Natural Language Engineering 2011;17:485-510.

[19] Zheng J, Chapman WW, Miller TA, Lin C, Crowley RS, Savova GK. A system for coreference resolution for the clinical narrative. J Am Med Inform Assoc. Published online first: January 31, 2012 doi: 10.1136/amiajnl-2011-000599

[20] Cia J, Strube M. Evaluation metrics for end-to-end coreference resolution systems. Proceedings of SIGDIAL 2010: The 11th Annual Meeting of the Special Interest Group on Discourse and Dialogue. 2010;28–36.

[21] Cunningham H, Maynard D, Bontcheva K. Text Processing with GATE (Version 6). Sheffield, UK: University of Sheffield Department of Computer Science, 15 April 2011.

[22] Cunningham H, Maynard D, Bontcheva K, Tablan V. GATE: A framework and graphical development environment for robust NLP tools and applications. Proceedings of the 40th

Anniversary Meeting of the Association for Computational Linguistics (ACL'02), Philadelphia, 2002.

[23] Wikipedia. Anatomical terms of location. Available from: http://en.wikipedia.org/wiki/Anatomical_terms_of_location [Accessed 31 January 2012].

[24] Wikipedia. Human anatomy. Available from: http://en.wikipedia.org/wiki/Human_anatomy [Accessed 31 January 2012].

[25] Rosse C, Mejino JV. A reference ontology for biomedical informatics: the Foundational Model of Anatomy. J Biomed Inform. 2003;36:478-500.

[26] Wikipedia. List of medical abbreviations. Available from: http://en.wikipedia.org/wiki/List_of_medical_abbreviations [Accessed 31 January 2012].

[27] Aronson AR. Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. Proc AMIA Symp. 2001;17-21.

[28] Gooch P, Roudsari A. A tool for enhancing MetaMap performance when annotating clinical guideline documents with UMLS concepts. IDAMAP Workshop at 13th Conference on Artificial Intelligence in Medicine (AIME'11).

[29] Lexical Systems Group. GSpell. Available from: http://lexsrv3.nlm.nih.gov/LexSysGroup/Projects/gSpell/current/index.html [Accessed 31 January 2012].

[30] Miller GA. WordNet: a lexical database for English. Communications of the ACM 1995;38(11):39-41.

[31] Chambers N, Jurafsky D. Unsupervised learning of narrative event chains. Proceedings of ACL-08: HLT, pp. 789-797. Columbus, OH: Association for Computational Linguistics, 2008.

[32] Cohen W, Ravikumar P, Fienberg S, Rivard K. SecondString: an open-source Java-based package of approximate string-matching techniques. Available from: http://secondstring.sourceforge.net/ [Accessed 31 January 2012].

[33] Cohen WW, Ravikumar P, Fienberg S. A comparison of string distance metrics for name-matching tasks. Proceedings of IIWeb. 2003;73-78.

[34] Winkler WE. String comparator metrics and enhanced decision rules in the Fellegi-Sunter model of record linkage. Proceedings of the Section on Survey Research Methods 1990;354–359.

[35] Monge AE, Elkan CP. The field matching problem: algorithms and applications. Proceedings of the Second International Conference on Knowledge Discovery and Data Mining. 1996;267-270.

[36] Savova GK, Chapman WW, Zheng J. Anaphoricity annotation guidelines for the clinical domain. J Am Med Inform Assoc. 2011;18(4)(online supplement 2).

[37] D'Avolio LW, Nguyen TM, Goryachev S, Fiore LD. Automated concept-level information extraction to reduce the need for custom software and rules development. J Am Med Inform Assoc. 2011;18(5):607-13.

9. Figure Captions

Figure 1. System architecture based around the GATE framework. ‘Thing’ refers to a non-Person mention such as AnatomicalTerm, Treatment, Test, Problem. Shaded areas represent components developed by the authors; unshaded areas represent existing components or external knowledge resources.

Figure 2. Filtering and traversal of mention pairs with pruning. Boxes represent mentions; vertical arrows represent iteration pointers; short horizontal arrows represent coreference pointers; shading represents mention features. (A) Document containing 2 classes of mention, differentiated by bold and dashed outlines. (B) Filtering of mentions of the same class (mentions 1, 2, 5, 7, 9, 10, 12) and start of iteration. (C) Identification of coreferent mentions 1 and 5. When the mention pointed to by the outer iterator matches a mention pointed to by the inner iterator, the features of antecedent (1) are cloned to the coreferent (5) (shown as shading in the figure), and the corefence pair is created. The antecedent is then pruned (D), and the outer pointer then moves to the coreferent (5) and the inner iterator increments (7) for the next iteration. (E) Identification of coreferent mentions 5 and 12 and addition of 12 to the coreference chain. (F) The inner iterator has completed, which closes the coreference chain, the previous coreference pair are pruned and the iterators reset. (G) Identification of coreferent mentions 2 and 9 and creation of a new coreference chain. (H) Antecedent pruning and outer iterator moves to coreferent. (I) Inner iterator completes, closing previous coreference chain, pruning of previous coreferent and iterators reset.

Fig. 1

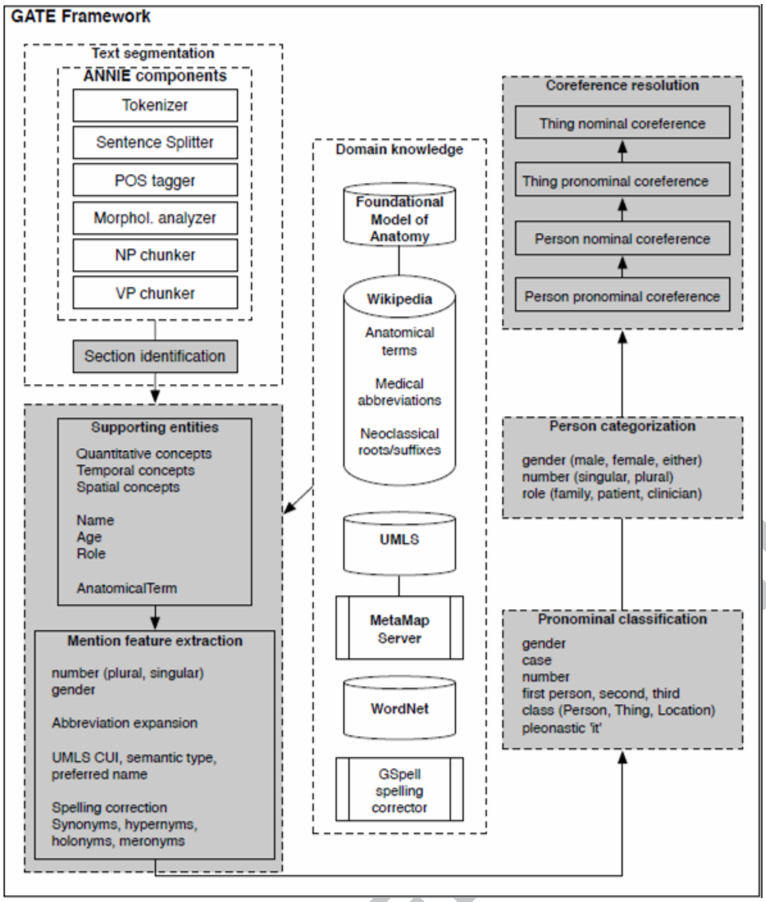
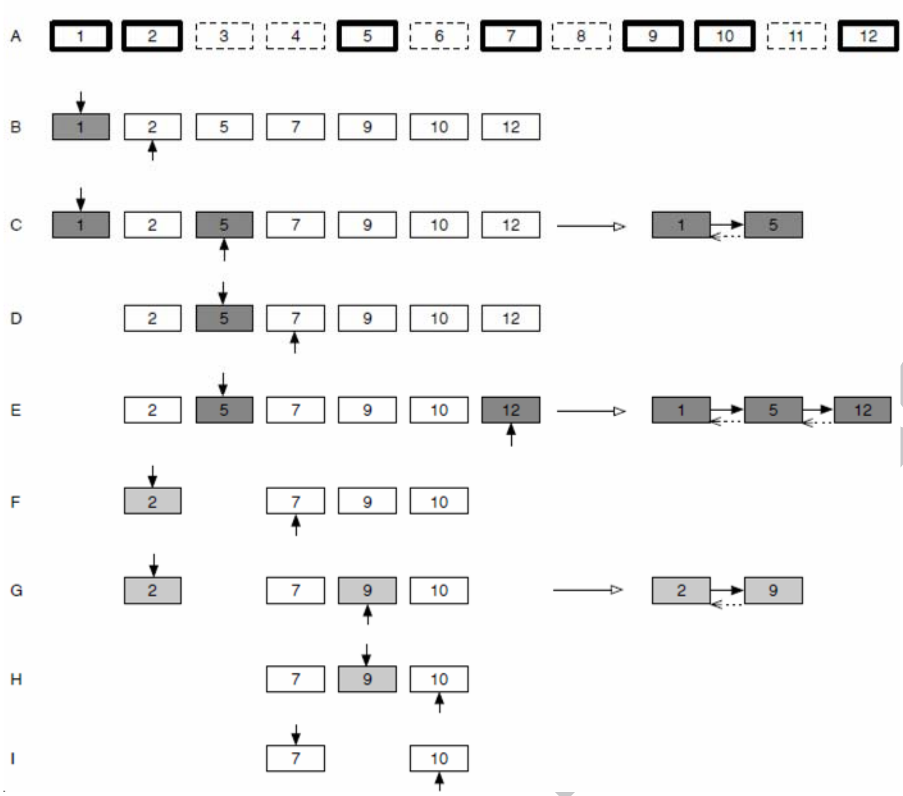
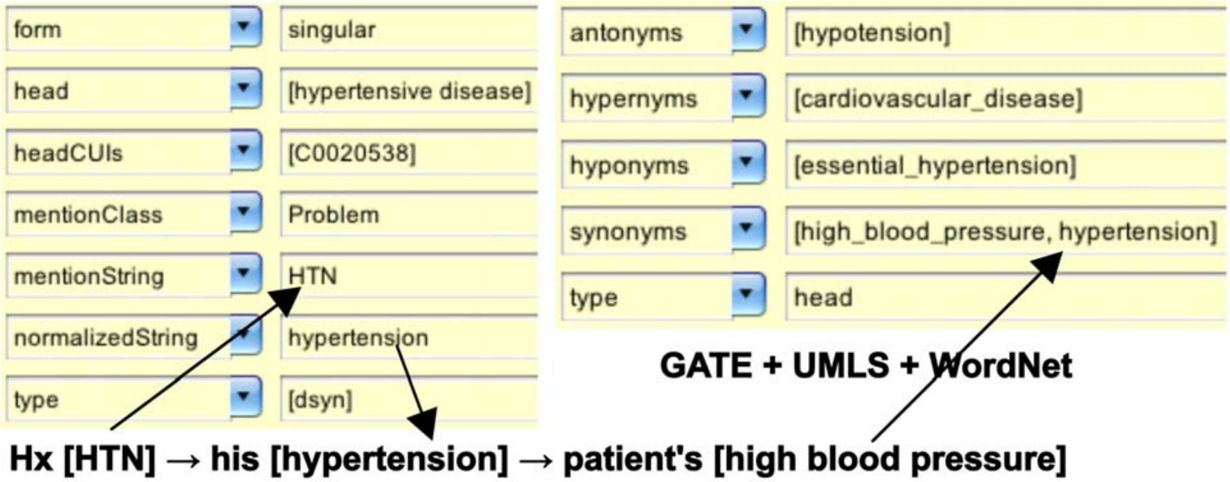

GATE Framework
Text segmentation
ANNIE components
Tokenizer
Sentence Splitter
POS tagger
Morphol. analyzer
NP chunker
VP chunker
Section identification
Supporting entities
Quantitative concepts
Temporal concepts
Spatial concepts
Name
Age
Role
AnatomicalTerm
Mention feature extraction
number (plural, singular)
gender
Abbreviation expansion
UMLS CUI, semantic type, preferred name
Spelling correction
Synonyms, hypernyms, holonyms, meronyms
Domain knowledge
Foundational Model of Anatomy
Wikipedia
Anatomical terms
Medical abbreviations
Neoclassical roots/suffixes
UMLS
MetaMap Server
WordNet
GSpell spelling corrector
Coreference resolution
Thing nominal coreference
Thing pronominal coreference
Person nominal coreference
Person pronominal coreference
Person categorization
gender (male, female, either)
number (singular, plural)
role (family, patient, clinician)
Pronominal classification
gender
case
number
first person, second, third
class (Person, Thing, Location)
pleonastic 'it'

Fig. 2


A
1
2
3
4
5
6
7
8
9
10
11
12
B
1
2
5
7
9
10
12
C
1
2
5
7
9
10
12
1
5
D
2
5
7
9
10
12
E
2
5
7
9
10
12
1
5
12
F
2
7
9
10
G
2
7
9
10
2
9
H
7
9
10
I
7
10

Graphical abstract



form	singular
head	[hypertensive disease]
headCUIs	[C0020538]
mentionClass	Problem
mentionString	HTN
normalizedString	hypertension
type	[dsyn]

antonyms	[hypotension]
hypernyms	[cardiovascular_disease]
hyponyms	[essential_hypertension]
synonyms	[high_blood_pressure, hypertension]
type	head

GATE + UMLS + WordNet
Hx [HTN] → his [hypertension] → patient's [high blood pressure]

Highlights

We present patterns for resolving coreference across a wide variety of clinical records

Our approach offers greatly increased performance over general purpose coreference tools

The system uses a number of open-source components that are available to download