Coreference resolution in clinical discharge summaries, progress notes, surgical and pathology reports: a unified lexical approach

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

Abstract

We developed a lexical rule-based system that uses a unified approach to resolving coreference across a wide variety of clinical records comprising discharge summaries, progress notes, pathology, radiology and surgical reports from two corpora (Ontology Development and Information Extraction (ODIE) and i2b2/VA) provided for the fifth i2b2/VA shared task. Taking the unweighted mean between 4 coreference metrics, validation of the system against the i2b2/VA corpus attained an overall F-score of 87.7% across all mention classes, with a maximum of 93.1% for coreference of persons, and a minimum of 77.2% for coreference of tests. For the ODIE corpus the overall F-score across all mention classes was 79.4%, with a maximum of 82.0% for coreference of persons and a minimum of 13.1% for coreference of diagnostic reagents. For the ODIE corpus our results are comparable to the mean reported inter-annotator agreement with the gold standard. We discuss the four categories of errors we identified, and how these might be addressed. The system uses a number of reusable modules and techniques that may be of benefit to the research community.

Introduction

Automated recognition of coreference, i.e. identity relations between noun phrases, and anaphoric relations, such as noun-phrase–pronoun pairs and part-whole relations, has become an increasingly important topic within the sphere of biomedical natural language processing. However, in a review of coreference methodologies, Zheng et al.[1] 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. [1]

Citing this lack of annotated corpora, Savova et al.[2] created a gold standard data set of clinical anaphoric relations – the Ontology Development and Information Extraction (ODIE) corpus. Their ongoing goal is to develop and evaluate tools and methodologies to resolve coreference in clinical texts. In this context, the focus of the fifth i2b2/VA challenge is on coreference resolution, and consists of three tasks:

- Task 1A: End-to-end identification, classification and intra-document coreference of concept mentions such as pronouns, people, procedures, diseases or syndromes, signs or symptoms, laboratory or test results, and anatomical sites, from the ODIE corpus[2]. The training set of 97 files consists of de-identified clinical notes and pathology reports from the Mayo Clinic, plus discharge summaries, progress notes, radiology reports, surgical pathology reports, and progress notes from the University of Pittsburgh Medical Center (UPMC).
- Task 1B: Intra-document coreference of concept mentions from the ODIE corpus, where mentions have already been identified and classified according to their UMLS semantic type.
- Task 1C: Intra-document coreference of concept mentions from the i2b2/VA corpus[3]. The training set of 492 files consists of de-identified discharge summaries from Partners HealthCare, Beth Israel Deaconess Medical Center, and UPMC. Mentions have already been identified and classified as Person, Problem, Treatment, Test and Pronoun (these are the ground truth concepts from the 2010 Challenge).

Inter-document coreference is out of scope for this challenge. In both the ODIE and i2b2/VA corpora, the guidelines used by human annotators to create the ground truth markables and coreference pairs were supplied to participants. The goal is to create a system that generates, for each input document, complete coreference chains for each discrete concept instance, in which the concept's class, and the text string and line/word offset of each co-referent markable are identified. For example:

```
c="right hip osteoarthritis" 21:0 21:2||c="advanced osteoarthritis of his right hip" 49:3 49:8||c="severe osteoarthritis of the right hip" 51:6 51:11||t="coref problem"
```

```
c="the patient" 22:0 22:1||c="she" 23:0 23:0||c="she" 24:0 24:0||c="her" 26:0
26:0||c="she" 27:0 27:0||c="she" 29:0 29:0||t="coref person"
```

In this paper, we describe our approach to Tasks 1B and 1C of this challenge and present cross-validation results for each data set within both training corpora.

Methods

The basis of our architecture is a rule-based system within the GATE[4] framework. The ground truth data sets provide line and word offsets for each markable, whereas GATE works with character offsets. We merged the concept markables, coreference pipeline file and source document for each record into a single XML file using the Knowtator conversion tool[5], and then wrote an XSLT transformation to convert this to the standoff XML annotation format used by GATE.

Using the i2b2/VA coreference annotation guidelines[3], the ODIE anaphoricity annotation guidelines[6], and a weighted random selection of 20 documents from both training corpora, we developed a set of general scenarios, applicable to both the i2b2/VA and ODIE corpora, that an automated system would need to consider:

1. Recognition of word synonyms, particularly for SignOrSymptom and Problem classes. For example, co-referring 'chills' with 'shivering' and 'inflammation' with 'swelling'.
2. Recognition of hypernyms and hyponyms, particularly for Problem, DiseaseOrSyndrome, Treatment and Procedure classes. For example, co-referencing 'staph bacteraemia' with 'the [infection]$_{\text{hypernym}}$', 'stereotactic biopsy' with 'the [procedure]$_{\text{hypernym}}$'.
3. Recognition of whole/part (meronym/holonym) relations, particularly for anatomical terms within mentions. For example, co-referencing 'the patient 's [head] wound laceration' with 'her [scalp]$_{\text{holonym}}$ laceration'.
4. Embedding of domain knowledge for both term equivalence, and processes related to or involved in the antecedent. For example, 'dyspnea' with 'shortness of breath' (term equivalence), 'CABG' with 'the revascularization' (process involved in the CABG procedure).
5. Automatic expansion and disambiguation of both standard and non-standard abbreviations. For example 'PT' in a Person mention might be expanded as 'patient', whereas in a Treatment mention it might be expanded as 'physiotherapy', and in a Test class as 'prothrombin time'.
6. Automatic correction/suggestion of likely misspellings, e.g. 'disciitis' (discitis), 'pian' (pain).
7. Handling agreement between the headwords of noun phrases where the antecedent is more specific than the coreferent, where all other contexts are equal. For example, 'intermittent right neck [swelling]$_{\text{head}}$' with 'the [swelling]$_{\text{head}}$'.
8. Rejection of coreference where the spatial concepts within each mention are different. For example, 'chronic [bilateral]$_{\text{spatial\_concept}}$ lower extremity swelling' should not be coreferenced with 'the [right]$_{\text{spatial\_concept}}$ lower extremity swelling'.
9. Rejection of coreference where the quantitative, spatial, temporal or anatomical context around each mention are different. For example: [2017-06-14 02:06AM]$_{\text{temporal\_context}}$: 'WBC - 9.4' vs. [2017-06-13 08:05PM] $_{\text{temporal\_context}}$: 'WBC - 9.4' and 'blood pressure' of [120/80]$_{\text{quantitative\_context}}$ vs. 'blood pressure' of [100/70]$_{\text{quantitative\_context}}$. Also 'simple atheroma' in the [aortic root]$_{\text{anatomical\_context}}$ vs. 'simple atheroma' in the [ascending aorta]$_{\text{anatomical\_context}}$.
10. Rejection of coreference between Problem, DiseaseOrSyndrome, Treatment and Procedure classes within sentences or sections of the document related to family history, or to other non-patient Person mentions.
11. Combinations of scenarios, e.g. scenario 1 and 7: '[ecchymosis]$_{\text{head}}$ of the right posterior thigh' vs 'right thigh [hematoma]$_{\text{head,synonym}}$'.
12. Classification of Person mentions in terms of their role within the document. This is necessary as the de-identification process leads, in some cases, to both patients and clinical staff being referred to by multiple

names, gender-neutral names, or random strings (e.g. WWWWW or **NAME[XXX YYY]). Does XXXX refer to the patient, or the physician treating them? Can we infer the gender of patient, family and clinician mentions, so that the likelihood of personal pronouns being related to one or the other can, in the absence of other linguistic cues, be calculated?

We implemented these scenarios as a set of linguistic rules in the Java Annotation Patterns Engine (JAPE) language in order to model the manual annotation process described in the guidelines and to generalize solutions to the sorts of problems encountered in the documents sampled. The rules were built on the output of an initial shallow parse and text segmentation pipeline (i.e. tokenization, sentence splitter, POS tagger, morphological analyzer, noun-phrase and verb-phrase chunking) using standard GATE ANNIE[4] components.

An overview of our system architecture is presented in Figure 1 and is described in detail below.

For scenarios 1-3 we developed a plugin that generates WordNet[7] annotations for given input mentions. We used the plugin to pass headwords, anatomical terms and general named entities (see below) to WordNet for discovery of lexical and semantic relations and stored the output as features within a WordNet annotation wrapped by the input mention.

For scenarios 4, 7 and 8 we used MetaMap[8] and the GATE mmserver integration plugin[9] 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 non-Person/Pronoun 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.

For scenario 5, we took a list of medical abbreviations from Wikipedia[10], and classified them according to their corresponding ground truth mention classes. A JAPE transducer was used to match abbreviations within mentions of the same class and store the expanded term as a feature on the mention.

For scenario 6, we developed a plugin using the GSpell library[11] and a dictionary derived from the `canonical.data` file from the LVG toolset[12]. This provides in-situ correction of misspelt non-Person/Pronoun 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.

For scenario 9 we 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. For anatomical terms not extracted by MetaMap term processing we extracted gazetteer lists of anatomical primitives (parts, spaces, locations, bones, muscles, organs) from Wikipedia[13,14] and the Foundational Model of Anatomy[15] and wrote JAPE rules to identify complete anatomical terms in the text via the logical combination of these primitives.

For scenario 10, the system splits source documents into sections and classifies each, based on the text of identifiable headings or paragraph content. Sections classified as being related to family history or historical lab data were then marked by the system as being excluded from coreference (depending on the mentions contained in these sections; for example, in a family history section Person and Pronoun mentions would still be eligible for coreference).

For scenario 12, we performed some preliminary analysis on the distribution of references to the subject of the report (i.e. the patient), and to personal pronouns within the set of all Person mentions in the ground truth concept files and coreference chains. Using regular expressions, we extracted Person-class coreference chains that contained the word "patient" or "pt" from the ground truth data. We also did this for personal pronouns (he, she, his, her etc) across all Person coreference chains and those within 'patient' coreference chains. Table 1 shows the distribution of patient and person pronoun references across the i2b2 and ODIE corpora.

Table 1. Distribution of Person and pronoun mentions in the ground truth markables and coreference chains.

	Person*	Personal pronouns	Patient*	Patient personal pronouns	Other pronouns
Coreference chains	19484	10313	14580	8879	1585
Markables	21127	10421	n.d.	n.d.	3862

* All mentions – includes personal pronouns. n.d. = not done - data not available

As shown in Table 1, 8879 out of 10313 (86.1%) personal pronoun mentions (that appeared in a coreference chain) are related to the patient. Of all coreferenced Person mentions, 14580 out of 19484 (74.8%) are in a patient coreference chain. The remaining mentions are related to members of the clinical team, to family/significant others, or to the person receiving the report. Therefore we considered Person mentions as being classified 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)

Person mentions were automatically classified according to the above types by the system using a set of JAPE rules and gazetteers of family relations (wife, daughter, brother etc), clinical roles (physician, doctor, nurse etc) and contextual cues (e.g. section heading content and gender identifiers). Given Table 1, non-pronoun Person mentions were classified as being related to the patient by default, unless the context suggested one of the other categories. Contextual cues (e.g. 'this is a 40-year old male') were then used to identify the gender of the patient. In the absence of cues, the document frequency of male and female pronouns were used to infer the patient's gender, given the prior probability (0.86) that a personal pronoun is related to the patient.

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. First-person pronouns 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 …'
- 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.

Pronoun classification

All non-personal pronouns are classified according to number (singular, plural), case (subjective, objective, possessive), person (first, second, third) and whether they represent pleonastic or anaphoric relations. We identify pleonastic 'it' and 'that' references using a set of general rules that look for temporal phrases ('It is Tuesday', 'it is 10pm'), 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 conjunction (e.g. 'It should be possible for …', 'It may be sensible to consider …'). Only anaphoric pronouns will participate in coreference.

Algorithms used for coreference chain resolution

Within the GATE framework we mark coreferent mentions by storing the internal annotation ID of the co-referring term on the coreferent, and create a back-reference by storing the ID of the coreferent on the coreferer. This prevents the creation of cycles in the coreference chain (by checking for the existence of forward or backward references prior to coreferencing pairs), and allows output to be deferred and performed by a separate component.

Coreference of Person mentions is simplified by the automatic classification process described above. We start with pronominal coreference. 'Who' pronouns are paired with the immediately preceding Person mention. Traversing in document order, pairs of nominal Person–third person-pronominal mentions are coreferenced if the genders, scope, role/type and number (singular or plural) agree.

To create linked chains, 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. The coreference chain is completed by matching nominal references, which now consist of nominal Person mentions and personal pronouns converted to nominal mentions by coreference.

Traversing in document order, pairs of 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 co-refer with sibling). Other Person mention pairs are coreferenced by evaluating the following, in order:

1. Exactly matching strings are coreferenced
2. Mentions with matching first names and surnames, where identifiable, are coreferenced
3. First person pronouns of global scope are coreferenced and linked to the primary clinician (usually the report's author)
4. Approximately matching strings are coreferenced. Using the SecondString Java library[16], and following Cohen et al.[17] we take the mean value of the Jaro-Winkler[18] and Monge-Elkan[19] 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.

All string matches are case-insensitive.

Coreference of clinical terms follows a similar approach as for Person mentions. Anaphoric pronouns are resolved against the most recent non-Person antecedent followed by the cloning of antecedent features to the anaphor. 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 scenarios 1-11 described above. Coreferencing is not attempted if either of the mention pair occurs in an excluded section (scenario 10) or if the contexts do not match (scenarios 8 and 9). A context match is made if there is a direct match between context features or there is a whole/part relation between the anatomical contexts of mention pairs.

1. If contexts match, or one of the pairs has a contextual feature and the other does not, coreferencing is provisionally attempted by attempting solutions to the remaining scenarios.
2. If there is an exact string match, the coreference is marked and iteration continues with the next mention pair.

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

1. 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);
2. There is intersection between sets of anatomical terms within each mention and between sets of UMLS semantic types for the headword/phrase;
3. The headwords and anatomical contexts match;
4. There is an approximate string match, as measured by the mean Jaro-Winkler/Monge-Elkan score within the defined threshold.

Although we developed a module for identifying negated and 'hedge-negated' mentions (e.g. 'surgery may not be appropriate'), we did not use this as the annotation guidelines were not clear on the scope of negated terms within the anaphoric relations to be identified, nor were we able to identify a pattern in the coreference of negated mentions in the ground truth documents that we sampled.

We performed 5 development iterations by running experiments against 5 weighted random selections of 10 records from both training corpora, analyzing the results in-situ using the corpus QA tools within GATE, and made adjustments to the process where errors could be generalized, for example, by making rules more (or less) specific and adjusting the scope of rules that were commonly misfiring. We did not make any document-specific changes to rules although we made use of bootstrapping to enhance the gazetteer lists of abbreviations.

Finally, we performed a full validation against both the ODIE and i2b2 corpora (589 documents) using the evaluation script supplied by the challenge organizers. The script reports evaluation results according to 4 coreference metrics implementations: B3, MUC, Blanc, and CEAF, as described by Cai and Strube[20].

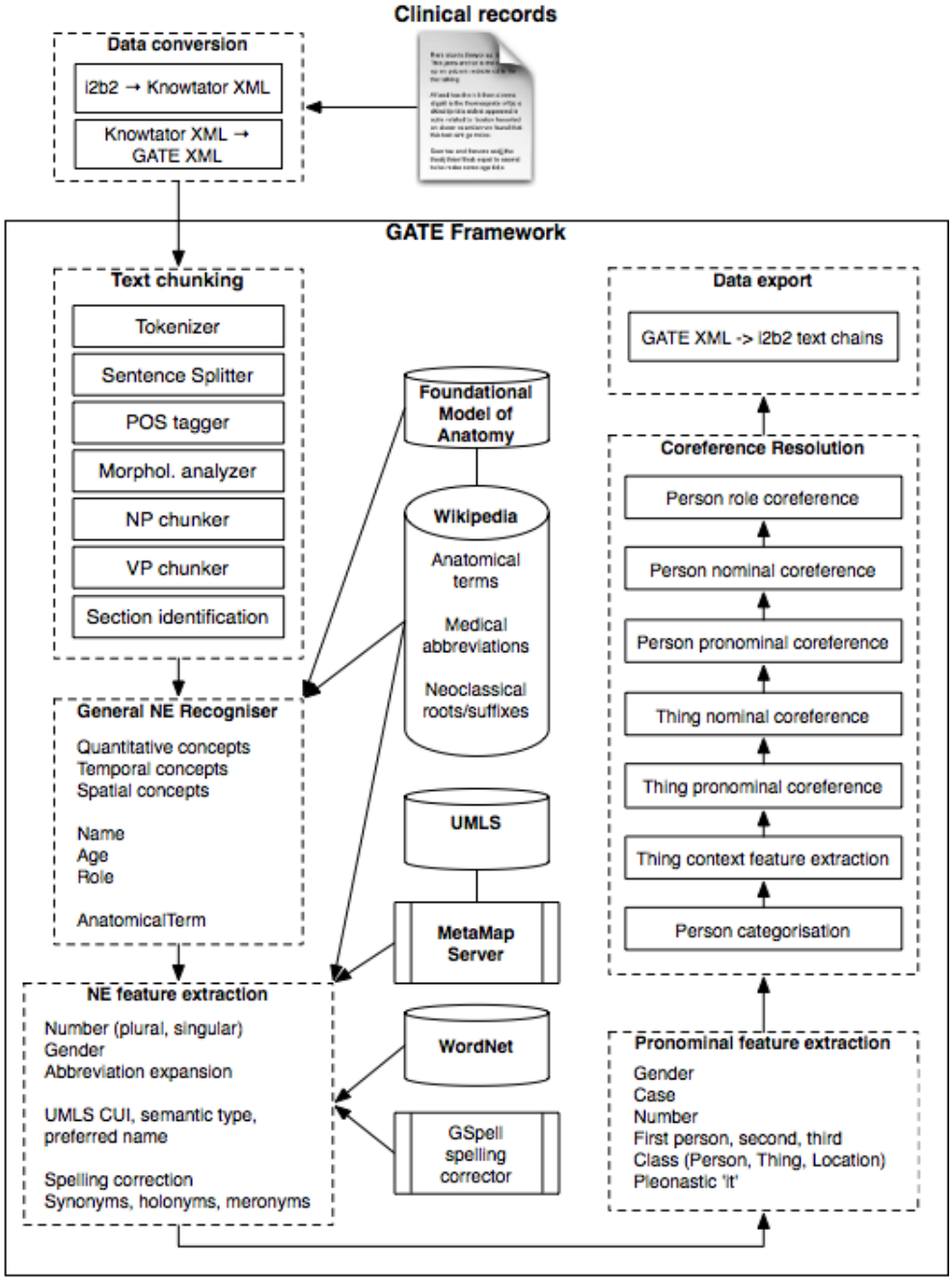

Figure 1. System architecture based around the GATE framework. 'Thing' refers to a non-Person mention.

Results

Validation results for records within the i2b2/VA training corpus are reported in Table 2, broken down by data set and mention class, and showing recall, precision, and *F*-measure scores for each metric for individual classes, and the micro-average *F*-measure across the B3, MUC and CEAF metrics for all classes. Table 3 shows results for the ODIE training corpus after correction of incorrect spans in the ground truth concept and coreference chain files (the full table of scores for the uncorrected UPMC data is not reproduced here; see Discussion).

Table 2. Coreference results for the i2b2/VA training corpus (492 documents; Task 1C)

	Metric												
	B3			MUC			Blanc			CEAF			Avg*
	Rec	*Prec*	*F*	*Rec*	*Prec*	*F*	*Rec*	*Prec*	*F*	*Rec*	*Prec*	*F*	
Beth Israel													
All classes	.966	.931	.948	.756	.878	.812	.978	.943	.960	.816	.896	.854	**.871**
Person	.956	.964	.960	.964	.970	.967	.987	.959	.973	.851	.880	.866	**.931**
Test	.975	.953	.964	.345	.598	.438	.771	.624	.671	.888	.940	.913	**.772**
Treatment	.964	.889	.925	.522	.779	.625	.837	.725	.770	.746	.870	.803	**.784**
Problem	.956	.918	.937	.637	.773	.698	.833	.766	.796	.812	.877	.843	**.826**
Partners Healthcare													
All classes	.971	.939	.955	.790	.878	.832	.979	.972	.976	.828	.890	.858	**.881**
Person	.904	.951	.927	.969	.953	.961	.986	.985	.986	.831	.769	.799	**.896**
Test	.964	.930	.947	.390	.662	.491	.784	.644	.691	.867	.933	.899	**.779**
Treatment	.961	.906	.933	.584	.797	.674	.861	.765	.806	.800	.895	.845	**.817**
Problem	.955	.917	.936	.595	.729	.655	.803	.769	.785	.826	.883	.853	**.815**
University of Pittsburgh Medical Center													
All classes	.972	.941	.956	.780	.882	.828	.965	.921	.942	.819	.894	.855	**.880**
Person	.881	.923	.902	.929	.923	.926	.971	.942	.956	.727	.709	.718	**.849**
Test	.965	.925	.945	.401	.724	.516	.833	.648	.705	.862	.939	.899	**.787**
Treatment	.972	.933	.952	.608	.813	.696	.872	.762	.807	.856	.922	.888	**.845**
Problem	.966	.923	.944	.556	.793	.654	.852	.691	.748	.802	.905	.851	**.816**

* Unweighted avg. of MUC, B3 and CEAF *F*-measures. Cross-class highest and lowest recall and precision scores highlighted.

Table 3. Coreference results for the ODIE training corpus (97 documents; Task 1B).

	Metric												
	B3			MUC			Blanc			CEAF			Avg*
	Rec	*Prec*	*F*	*Rec*	*Prec*	*F*	*Rec*	*Prec*	*F*	*Rec*	*Prec*	*F*	
Mayo													
All classes	.854	.933	.892	.845	.709	.771	.888	.959	.920	.803	.589	.680	**.781**
People	.830	.952	.887	.960	.918	.939	.929	.974	.950	.764	.541	.633	**.820**
Disease	.819	.895	.855	.632	.434	.514	.612	.776	.658	.792	.577	.667	**.679**
Symptom	.901	.926	.913	.662	.544	.597	.619	.707	.605	.817	.810	.814	**.775**
Anat. Site	.844	.920	.880	.678	.500	.576	.669	.826	.723	.783	.603	.681	**.712**
Reagent†	.174	1.00	.296	.000	.000	.000	.500	.397	.329	.306	.057	.096	**.131**
Organ Fn. †	.250	1.00	.400	.000	.000	.000	.500	-.500	.000	.667	.667	.667	**.356**
Lab. Result	.778	.875	.824	1.00	.500	.667	.667	.800	.625	.933	.700	.800	**.764**
Procedure	.859	.928	.892	.641	.465	.539	.675	.730	.699	.839	.739	.786	**.739**
University of Pittsburgh Medical Center													
All classes	.893	.918	.905	.850	.820	.835	.946	.944	.945	.721	.644	.680	**.807**
People	.860	.902	.880	.928	.923	.925	.961	.948	.954	.529	.521	.525	**.777**
Disease	.858	.897	.877	.728	.665	.695	.788	.844	.813	.807	.695	.746	**.773**
Symptom	.949	.906	.927	.609	.739	.668	.759	.783	.771	.768	.854	.809	**.801**
Anat. Site	.770	.873	.818	.739	.628	.679	.641	.820	.693	.536	.373	.440	**.646**
Reagent†, ††	-	-	-	-	-	-	-	-	-	-	-	-	**-**
Organ Fn.	.593	.770	.670	.714	.556	.625	.533	.561	.489	.704	.302	.422	**.572**
Lab. Result††	1.00	.957	.978	.000	.000	.000	.498	.500	.499	.895	.971	.932	**.637**
Procedure	.736	.880	.802	.803	.560	.659	.683	.868	.743	.769	.573	.657	**.706**

* Unweighted avg. of MUC, B3 and CEAF *F*-measures. † 0 system results †† 0 ground truth results; treat scores with caution.

The highest and lowest cross-class precision and recall scores for non-null system results in each data set are shaded in both tables. High precision and/or recall scores in cases of null results in either the system output or ground truth key set do not appear to be meaningful. As expected, the B3 scores are generally higher than the MUC scores due to the leniency of B3 towards twinless coreferences[20] and broken coreference chains. However, MUC more correctly reports zero precision and recall in cases where the system gives a null result when coreference chains exist in the key set, and vice versa.

Discussion

We identified 4 categories of error in our system validation:

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. *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 physican names with their clinical role.
4. *Data conversion errors*: We noted that the use of line and word offsets in the input and output data can be a source of ambiguity and error as they are dependent on both the choice of line termination character(s) and choice of tokenization algorithm. Internally, our system uses character offsets, which requires a fairly convoluted input/output conversion process that was also a source of error. This has now been fixed, although unfortunately not before submission of our results for the withheld test evaluation data.

Instances of incorrect reporting of word offsets in the ground truth markables, caused by leading white space at the start of a line being incorrectly counted as a word, initially led to low scores for the ODIE UPMC corpus, particularly for the surgical/pathology reports. We corrected the word offsets in the ground truth data and ran the evaluation again; this led to an increase in the overall *F*-measure from 61.2% to the 80.7% reported here.

For the i2b2/VA corpus, our system performed best at coreferencing Person mentions, for which there was generally good agreement between evaluation metrics across all 3 data sets. However, it performed least well at coreferencing Tests – the particularly low recall, as measured by MUC for this class, reflected the system's over-zealous exclusion of all test results within lab data sections.

Relative system performance between classes was similar for the ODIE corpus. Here, 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'.

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, which may partially explain the weaker results in comparison to those for i2b2/VA. However, the ODIE results are comparable to – or, for the UPMC data set, better than – the mean IAA reported between annotators and the gold standard for each dataset[2] (81.29% and 69.45%, compared to our system performance of 78.1% and 80.7%, respectively).

Our system performance may be improved in the following ways. It could be modified to make use of the ground truth line/word mention spans directly, rather than generating these from the internal character offsets of each mention. We could implement a more specific set of rules for identifying pleonastic references, using patterns described by Dimitrov et al.[21] We could make use of full dependency parsing to generate more precise contextual cue features. Currently, the coreferencing step relies on lexical rules to process the mention features extracted by the system. These could be enhanced with the use of machine learning algorithms. The name matching process could be modified, for example removing the Jaro-Winkler/Monge-Elkan string similarity step (which tended to give false

positives for de-identified name fragments) and instead checking for a surname-only match, although this would be dependent on the de-identification process used.

Conclusion

We have developed a lexical rule-based system that uses a common approach to resolving coreference across a wide variety of clinical records comprising discharge summaries, progress notes, pathology, radiology and surgical reports from a variety of sources. The system uses a number of reusable modules and techniques that may be of benefit to the research community. Overall the system is fairly strong at coreference resolution of persons, treatments/procedures and clinical problems/symptoms/diseases, but is weaker at resolving coreference of anatomical and physiological terms, laboratory tests and results. We have validated the system against two training corpora, and aim to report final results for its performance against the withheld test corpora in due course.

Acknowledgements

Deidentified clinical records used in this research were provided by the i2b2 National Center for Biomedical Computing funded by U54LM008748 and were originally prepared for the Shared Tasks for Challenges in NLP for Clinical Data organized by Dr. Ozlem Uzuner, i2b2 and SUNY. Part of the data has been produced by the ODIE grant (R01 CA127979, NCI/NCBI, PI Crowley) and contributed to the i2b2/VA 2011 challenge under SHARP 4 (U01 SHARP 4, ONC, PI Chute). The author 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.

References

1. 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; in press.
2. 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.
3. Uzuner O. 2011 i2b2/VA co-reference annotation guidelines for the clinical domain. Available from: https://www.i2b2.org/NLP/Coreference/assets/CoreferenceGuidelines.pdf [Accessed 31 August 2011]
4. 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.
5. OGren PV. Knowtator: A plug-in for creating training and evaluation data sets for biomedical natural language systems. 9th International Protégé Conference. 2006.
6. Savova GK, Chapman WW, Zheng J. Anaphoricity Annotation Guidelines for the Clinical Domain. J Am Med Inform Assoc. 2011;18(4)(online supplement 2).
7. Miller GA. WordNet: a lexical database for English. Communications of the ACM 1995;38(11):39-41.
8. Aronson AR. Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. Proc AMIA Symp. 2001;17-21.
9. 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)
10. Wikipedia. List of medical abbreviations. Available from: http://en.wikipedia.org/wiki/List_of_medical_abbreviations [Accessed 31 August 2011]
11. Lexical Systems Group. GSpell. Available from: http://lexsrv3.nlm.nih.gov/LexSysGroup/Projects/gSpell/current/index.html [Accessed 31 August 2011]
12. Lexical Systems Group. The SPECIALIST lexical tools. Available from: http://lexsrv3.nlm.nih.gov/Specialist/Summary/lexicalTools.html [Accessed 31 August 2011]
13. Wikipedia. Anatomical terms of location. Available from: http://en.wikipedia.org/wiki/Anatomical_terms_of_location [Accessed 31 August 2011]
14. Wikipedia. Human anatomy. Available from: http://en.wikipedia.org/wiki/Human_anatomy [Accessed 31 August 2011]
15. Rosse C, Mejino JV. A reference ontology for biomedical informatics: the Foundational Model of Anatomy. J Biomed Inform. 2003;36:478-500.
16. 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 August 2011]

17. Cohen WW, Ravikumar P, Fienberg S. A Comparison of String Distance Metrics for Name-Matching Tasks. Proceedings of IIWeb. 2003;73-78.
18. 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.
19. 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.
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. Dimitrov M, Bontcheva K, Cunningham H, Maynard D. A light-weight approach to coreference resolution for named entities in text. Proceedings of the Fourth Discourse Anaphora and Anaphor Resolution Colloquium (DAARC), Lisbon. 2002.