A Search Based Method for Clinical Text Coreference Resolution

Demetrios G. Glinos, JD, PhD
Advanced Text Analytics, LLC, Orlando, Florida, USA

Abstract

This paper describes a novel method and system developed to address the coreference task of the 2011 i2b2 NLP Challenge, which involved analyzing clinical texts of several types and identifying the coreferential mentions within them. The method uses an open source search library component to do the heavy lifting for string matching, allowing a lightweight rule-based processing engine to cluster the concept mentions in an almost entirely annotation-agnostic manner. The system performed well against the challenge test data set, which included documents from multiple clinical sources that were annotated according to both i2b2 and ODIE guidelines, and achieved the best overall score against the ODIE test data. The results show promise for further investigation of the approach.

Introduction

Coreference resolution is the process of identifying which concept mentions in a document refer to the same discourse entity. It includes, for example, identifying that "severe three-vessel disease" in one paragraph of a document refers to the same concept as "coronary artery disease" in the preceding paragraph. It is an essential task in the process of converting unstructured text, as are many clinically relevant documents even in their electronic form, into structured form so that it can be queried, summarized, translated, analyzed, or otherwise made useful to automated processing systems. Unstructured text is rife with myriad ways in which the same discourse entity can be referenced in the same document, often in the same paragraph, and sometimes even within the same sentence. A person, for example, can be referred to by: (a) his or her full name, possibly with or without title, suffix, or degree; (b) a nickname or a subset of the full name, such as the first or last name, again with or without title, etc.; (c) by a nominal expression, such as "the patient" or "the patient's spouse"; or (d) by a pronominal such as "he", "her", or even "who". Clinical text in particular often includes abbreviations and shorthand expressions, such as "pt" for "patient", which are useful for saving time in the hectic clinical environment. Such text also often contains "noise" in the form of incomplete sentences, faulty grammar, and otherwise well-formed clauses embedded within the confines of semi-structured document templates.

To address the problem in the clinical domain, and also to encourage cross fertilization with natural language processing techniques developed for open-domain applications, Informatics for Integrating Biology and the Bedside (i2b2)[1] sponsored the 2011 NLP Shared Task ("2011 NLP Challenge"). The topic for Track 1 of the Task was coreference resolution and involved manually annotated de-identified data from: Partners HealthCare, Beth Israel Deaconess Medical Center, the University of Pittsburgh, and the Mayo Clinic. The data included a variety of medical record types, including discharge summaries, pathology reports, radiology reports, and surgical pathology reports. Part of the data was annotated according to i2b2/VA guidelines[2], which include markables for "problem", "test", "treatment", and "person". A "pronoun" type was also used for relative pronouns, such as "who", "which", "this", "these", "it", and "that". Another part of the data was annotated according to ODIE guidelines[3], which include markables for "people", "procedure", "test", "diseaseorsyndrome", "signorsymptom", "anatomicalsite", "laboratoryortestresult", "indicatorreagentdiagnosticaid", and "organortissuefunction". ODIE and i2b2/VA documents were also tokenized differently: i2b2/VA documents contained one "sentence" per line in the file, with punctuation separated from the adjoining words, while ODIE data sentences spanned multiple lines and did not have punctuation separated from text. The training set included 492 i2b2/VA documents and 97 ODIE documents. The Test set included 322 i2b2/VA documents and 66 ODIE documents. Gold standard chain files were furnished for the training data.

The coreference track consisted of three subtasks. Track 1A involved a test against ODIE data in which only un-annotated source text documents were furnished. For this subtask, participants first extracted the concept mentions and then grouped them into coreferential "chains" or clusters. Track 1B also involved ODIE data, but provided files identifying the concept mentions and their spans within the text for each raw text document. For this subtask, participants performed only the clustering needed for coreference resolution. Task 1C was similar to Task 1B, except that it ran against i2b2/VA data. We participated only in Tasks 1B and 1C and report results against the test data set in this paper.

For the past decade, coreference resolution research has been dominated by supervised machine learning methods, in which a statistical classifier is learned based on annotated training examples, and then is applied to new instances of the same type of data. Early efforts used the "mention-pair" model, which focused on mining the associations among pairs of noun phrases (NP), as in [4]. More recent efforts have focused on the "entity-mention" model, in which an NP is associated with an already existing coreferential cluster, the members of which provide a richer feature set for training, with improved results. An example of this type is [5], which involved a classifier trained on a set of 70 annotated MEDLINE abstracts from the GENIA data set, and tested on a further 30 abstracts, with promising results. Current efforts have focused on variations on the entity-mention model, such as [6], which involved using inductive logic programming to represent a fuller set of relations among mentions in a coreferential cluster, and also [7], which involved unsupervised learning using Bayesian methods, and [8], which uses a generative model for unsupervised learning.

Document retrieval is a common activity in contemporary life. Search engines and applications possess well-known powerful capabilities to retrieve documents and information using search string inputs. Many of these systems are built around open source software search library components whose merits have been established in the field of large-scale "real world" document and database applications. One such open source search library is Lucene[9], from the Apache Software Foundation. It is written entirely in Java, is stable, has an active developer network, and is used in a large number of well-known publicly accessible sites, as may be gleaned from its "powered by" page[10].

In this paper, we report on what we believe to be the successful first use of such open source search technology for coreference resolution. In the next section, we present the details of our motivation and method, followed by a section presenting the results against the NLP Challenge test data, including some additional ablation tests to assess the relative impacts of the search and knowledge based components. We conclude with a section that discusses the limitations of the method and suggests avenues for further investigation.

Method

Our approach for this effort was to use the data to drive the development of the method. Consider, for example, the sample discharge summary shown in Figure 1, taken from the NLP Challenge training data set:

```
**INSTITUTION
Discharge Summary
Name :
**NAME[AAA , BBB]
Acct # :
**ID-NUM
MRN :
**ID-NUM
Admitted :
**DATE[Dec 11 2007]
Discharged :
**DATE[Feb 20 2008]
Dict :
**NAME[XXX , M. WWW]
Attend :
**NAME[ZZZ , YYY]
DEATH SUMMARY :
Mr. **NAME[AAA] is a **AGE[in 60s]- year - old white male with a history of metastatic
cholangiocarcinoma with recurrent aspiration and sepsis who underwent palliative measures
for respiratory distress .
The patient was pronounced dead at 1:41 a.m.
Family was notified .
Autopsy was declined .
_____
**NAME[M. WWW XXX] , MD
HS Job # 306548 / 38345 / 35066
SHY # **ID-NUM
D :
**DATE[Feb 20 2008] 02:13
T :
**DATE[Feb 21 2008] 05:12
**CARBON-COPY
```

Figure 1. Sample NLP Challenge source document.

Although shorter than many other documents in the training data set, this document possesses a number of the features that drove the development of our approach when considered in the context of its concept mentions and their spans. Figure 2 shows the gold standard chain file for the document in Figure 1, reporting the clustered concept mentions, from which we observe that the de-identified patient is referred to as "**name[aaa, bbb]" in line 4 and as "mr. **name[aaa]" in line 18. Similarly, the dictating physician is referred to as "**name[xxx , m. www]" in line 14, but as "**name[m. www xxx] , md" in line 23. Although a regular expression or a class method could handle associating these concept pairs, we believed that there are in general too many permutations and combinations to consider, so that attempting to enumerate them in this manner would be ultimately unsuccessful. Instead, based on our prior experience in using Lucene, we felt that the search library could handle making these kinds of associations.

```
c="**name[aaa , bbb]" 4:0 4:2
        ||c="mr. **name[aaa]" 18:0 18:1
        ||c="who" 18:22 18:22
        ||c="the patient" 19:0 19:1||t="coref person"
c="dict" 13:0 13:0
        ||c="**name[xxx , m. www]" 14:0 14:3
        ||c="**name[m. www xxx] , md" 23:0 23:4||t="coref person"
c="attend" 15:0 15:0
        ||c="**name[zzz , yyy]" 16:0 16:2||t="coref person"
```

Figure 2. Gold standard chain file for sample NLP Challenge source document.

Examining the chain file further, we observe that "mr. **name[aaa]" is also clustered with "who" and "the patient". Making these associations, we felt, was beyond Lucene's capabilities, for two reasons: (a) the terms are not at all similar at the lexical level; and (b) these associations require knowledge of the surrounding discourse context. Therefore, in our view, a second system component was needed, and we chose to proceed with a rule-based component as the most straightforward for combining the initial clusters identified by the search component.

These choices raised two additional design questions: (a) how to use the search component to make the desired first-level associations; and (b) how to preserve context so that the subsequent associations above could be made by the rule-based component. To answer the first question, we noted that in a typical search application it is the raw documents that are indexed and retrieved, as in a typical Google search over the Internet, but in our case what we wished to retrieve at the first level was the set of concept mentions that were similar to a given mention. We therefore chose to create an index using the concept mentions themselves as documents, and to use Lucene's fuzzy search feature to retrieve the similar concept mentions. To answer the second, the context that would need to be examined by the rule-based component would be the source document itself. However, since we needed essentially random access into the file in order to examine the syntactic features surrounding concept mentions, we chose to store each line of the document separately, leaving it to the context examination methods to take into account that a line was a sentence for i2b2/VA documents, but not for ODIE documents.

Requirements were also derived from examining additional associations discovered in the gold standard chain files. For example, to associate "Postpolio" with "Postpoliomyelitis", we identified the need for examining lexical variants of search terms. For this purpose, we selected the open source Lexical Tools[11], available from the National Library of Medicine's Lexical Systems Group. This toolkit includes a lexical variant generator that uses the NLM's SPECIALIST lexicon of biomedical and general English. And to associate "resection" with "surgery", we identified the need to examine a search term's hypernyms. For this purpose, we selected WordNet[12], which though not open source is nonetheless freely available from Princeton University. WordNet is a large lexical database of common contemporary English nouns, verbs, adjectives and adverbs, whose terms are grouped into conceptually synonymous classes called "synsets", which are in turn related semantically and lexically in various ways, including hypernymy. Though not including medical terminology *per se*, WordNet's hypernym and synonym relations are particularly relevant to the coreference problem.

Based on the considerations above, our method is depicted in the processing flow shown in Figure 3. Processing is performed for each text and concept document pair before proceeding to the next pair. For each document pair, the first step is to create a document database where each entry is a line of the source document and also to create a separate search index of the concept mention text strings.

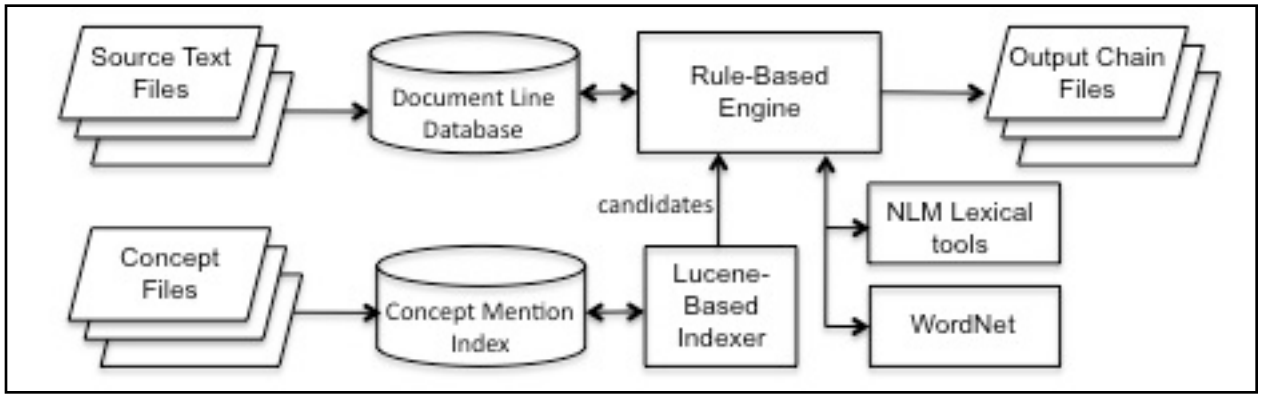

Figure 3. Coreference resolution processing flow.

The next step is to cycle through each concept mention in the document and to perform a search to obtain a set of similar candidate concept mentions. The search is performed using Lucene's "fuzzy query" capability[13], which implements the Levenshtein distance (also known as "edit distance") algorithm. This metric computes the distance between two strings as the minimum number of insertion, deletion, or character substitution operations necessary to convert one string into another. Lucene further applies a threshold factor, computed as the distance divided by the length of the shorter of the two strings. The metric is computed across all concept terms in the index and those that exceed the threshold are reported.

The candidates are passed to the rule-based processing engine, which determines whether or not to accept candidate concepts as coreferent. After first excluding pairs that involve "that", "which", "who", "this", "that", "these", and "it", the engine separates pairs in which both terms involve "person" or "people" markables. Such pairs that are named entities are tested for compatibility with respect to gender, number, and titles, the last so that a "Dr." or "MD" is not matched with a "Mr." or "Ms.". For those that pass these tests, their normalized values are obtained using the NLM Lexical tools. The normalizations exclude punctuation, articles, pronouns, and prepositions. If the normalizations have any terms in common, then the original terms are deemed accepted. For candidate matches involving two pronominals, only an exact match is accepted at this stage. For all other candidate pairs, they are first checked for compatibility with respect to number, where that can be determined, and then their normalized terms are compared for common values as for named values. Once a concept mention has been assigned to a coreference cluster in this manner, it is not reconsidered for assignment to any other cluster.

Following this initial similarity determination, the rule-based engine performs three additional steps. The first is to revisit each concept mention in the document that has not been assigned to a coreference cluster and to examine it for similarity with respect to every concept cluster that has been created. The idea behind this step is to consider pairings without regard to Lucene's search threshold. The same criteria for similarity as described above are applied, and any new associations are implemented as they are found.

The second additional step is to apply a set of heuristic rules for associating concepts in the clinical domain. These rules are applied both to unassigned concepts ("singletons") and to existing clusters. These rules include:

1. Attaching headings: Some of the clinical documents identify headings as concept mentions. Where a heading is immediately followed by a concept of the same type, they are associated. And example is "Attending" in the heading "Attending Name :", followed by a physician's name on the next line. Heading associations can be made for any annotation type, not just for "person" or "people" types, provided the types for both entities agree.

2. Departments as Persons: It is common in clinical records to find that department names are used to refer to persons from a department, as in, for example, "Cardiology reported that...". Where a department name has been marked as a person concept mention, the mention is annotated to reflect that it is a named entity, specifically a physician, but given the "plural" number. The current departments implemented include: "cardiology", "dermatology", "er", "hematology", "immunology", "orthopedics", "path", "pathology", "physical therapy", "radiology", "surg", "surgery", "surgpath".

3. Identifying Doctors: Doctors not identified as above are identified through their titles (e.g., "Dr.") or degree (e.g., "M.D."), and also if they preceded by a heading indicating a physician role. The headings currently

implemented include: "attending", "attend", "author", "dictator", "dict", "dictated", "doctor", "pcp", "provider", "physician", "primary", "consult", "signed", "signer", and "service".

The third additional step is to apply a set of heuristic rules for merging coreference clusters. These rules are as follows:

A. Attaching Nominals: For nominal entities, such as "the patient", this rule first checks whether it occurs as a comma-separated apposition either before or after another concept of the same type. If so, then the two are associated. If not, then if it is a "person" or "people" type, the text is searched backwards from the current position to find the nearest concept of the same type that matches with respect to gender, number, doctor status (so that patient nominals cannot be associated with doctors), and even family member (so that a spouse is not matched with the patient). If no concept mention matches these criteria, then the concept or cluster is not associated.

B. Attaching Pronominals: Personal pronouns are separated into two categories. First person pronouns ("I", "we", etc.) are checked for attachment to named entities that are marked as physicians. If the dictator of the document is identified, then that association is made. If not, then the document is searched from the bottom up for a physician who is not the signer, since many, but not all, of the clinical documents include person markables for the physicians who dictated or signed the document. For non-first person pronouns, such as "he" or "she", a search is made from the position of the pronominal going backwards through the document for a named entity of the same type that also matches in gender, number, who is not a doctor, and who is similar in the same manner as candidates received initially from the indexer.

C. Attaching Relatives: Concept mentions for the term "who", however marked, are checked against the immediately preceding concept mention; if that mention is also a person, then the two are associated. Relative pronouns "that" and "which" are matched to the first encountered concept of the same type that is encountered in a backward search through the document starting from the current position. For "this" or "these", the text following the concept mention is examined until either a noun, verb, or punctuation is encountered. If a noun is found, then the position of that noun is noted and the concept list is examined for mentions that include that word. If such a concept is found, then the relative pronoun is associated with that mention. If a verb is found, then the relative pronoun is attached to the next occurring non-relative that is encountered in a search proceeding forward from the current position. If punctuation is found before either a noun or verb, then the relative pronoun is not associated by this rule.

D. Merging Pure Pronouns: This rule operates on non-relative pronouns, not on relative pronouns that are annotated as "pronoun" in the i2b2/VA test data. If a coreferential cluster contains only pronominals, then this rule will merge it with another cluster that contains pronominals that agree in gender and number. For example, this rule will merge a chain of "he" references with a chain of "him" references.

E. Merging Pure Patient Chain: This rule will operate to attach a chain containing only nominal references containing the word "patient" to a chain that contains only pronominal terms, provided there is only one such chain.

F. Merging Pure Disease Chain: This rule first finds all chains of type "problem" or "diseaseorsyndrome". If there are only two and one of them contains only nominal references containing the word "disease", then the two chains are merged.

Once all rules are applied once, the coreference chains that were created are output to a file, the database and index are cleared, and the next text and concept document pair is processed. The system does not report single-member chains for unassigned concept mentions (“singletons”).

Results

The method described was implemented and tested formally against both i2b2/VA (Task 1C) and ODIE (Task 1B) test data sets in the 2011 NLP Challenge. We also subsequently performed ablation tests to ascertain the relative performance contributions of the search-based, rule set, and knowledge-source components. All runs were evaluated using the evaluation script furnished by i2b2. The script performs micro-averaged recall, precision, and F-measure calculations according to several metrics. We report here the results according to the BCUBED[14], MUC[15], and CEAF[16] measures, as these are the metrics that were averaged to compute the official scores for Tracks 1B and 1C.

Table 1 presents the results of the full (unablated) system against the i2b2 and ODIE test data sets. Despite the differences in annotation schemes, the overall performance was quite similar for both data sets. For the i2b2 data, average precision was high at 87.3%, but average recall was modest at 80.0%. As a result, the average F-measure was 83.3%, which was slightly below the median of all systems that participated in Task 1C.

Table 1. Full system overall coreference resolution performance.

	CEAF	Bcubed	MUC	Averages
I2b2 Data				
Recall	0.736	0.962	0.703	0.800
Precision	0.855	0.913	0.852	0.873
F-Measure	0.791	0.937	0.771	0.833
ODIE Data				
Recall	0.752	0.899	0.876	0.842
Precision	0.680	0.936	0.825	0.814
F-Measure	0.714	0.917	0.850	0.827

By contrast, against the ODIE data, the system had higher average recall (84.2%), lower average precision (81.4%), and lower average F-measure (82.7%). Nevertheless, the average F-measure of 82.7% achieved by our system against the ODIE data was the highest of all systems that participated in Task 1B.

Table 2 shows how the system performed against the different categories of i2b2 data. From the table, it is clear that the overall score was significantly adversely affected by the anomalously low MUC scores for "test" markables. Investigation into the causes of this phenomenon revealed that the system tended to be overly zealous in clustering concepts, in one instance for example grouping "blood cultures", and "blood glucose checks" with "blood work". A better similarity filtering rule, perhaps based on deeper domain knowledge, with improved results for "test" markables, and perhaps also for "treatment" and "problem" markables, would no doubt have improved the overall score considerably. By contrast, performance on "person" markables, with which much of the clustering rules were concerned, was the best among the i2b2 categories, scoring a quite respectable average F-measure of 84.9%.

Table 2. Full system performance against i2b2/VA data by markable type.

	Person	Test	Treatment	Problem
CEAF				
Recall	0.640	0.644	0.795	0.764
Precision	0.790	0.889	0.877	0.880
F-Measure	0.707	0.747	0.834	0.818
Bcubed				
Recall	0.925	0.972	0.949	0.952
Precision	0.899	0.871	0.923	0.915
F-Measure	0.911	0.918	0.936	0.933
MUC				
Recall	0.909	0.169	0.614	0.557
Precision	0.948	0.555	0.726	0.710
F-Measure	0.928	0.259	0.665	0.624
Averages				
Recall	0.825	0.595	0.786	0.758
Precision	0.879	0.772	0.842	0.835
F-Measure	0.849	0.641	0.812	0.792

Table 3 reports system performance against the different ODIE markables that occurred in the test data set. Some ODIE markable types are not represented in the table because they did not occur in the test set, which, as previously reported, contained only 66 documents. The perfect scores for the "organortissuefunction" type were due to the fact that there were only six mentions of this type in three documents and the system correctly clustered them. The zero value MUC scores for the "laboratoryortestresult" markable were caused by the fact that there were only 19 mentions of this type in the data set and all of them were singletons, that is, none of them appeared in coreferential chains. Since the MUC measure is based on links, the singletons were not counted, resulting in the zero scores.

Interestingly, the scores for the "people" markable were not markedly better than for the other types, as was the case for the i2b2 data. We have not performed a systematic investigation into the cause; however, we believe that the

reason is due to the relatively more terse nature of the document types that were annotated using the ODIE scheme, such as the pathology, radiology, and surgpath report types, which may have spoofed the rules that associated pronouns (likely referring to the patient) with named persons (more likely in this data to represent the physician).

Table 3. Full system performance against ODIE data by markable type.

	Anatomical Site	Organ or Tissue Function	Sign or Symptom	Laboratory or Test Result	People	Procedure	Disease or Syndrome
CEAF							
Recall	0.719	1.000	0.788	0.614	0.545	0.816	0.776
Precision	0.617	1.000	0.880	0.897	0.545	0.751	0.736
F-Measure	0.664	1.000	0.832	0.729	0.545	0.782	0.756
Bcubed							
Recall	0.880	1.000	0.955	1.000	0.864	0.863	0.891
Precision	0.921	1.000	0.946	0.862	0.880	0.956	0.934
F-Measure	0.900	1.000	0.951	0.926	0.872	0.907	0.912
MUC							
Recall	0.787	1.000	0.677	0.000	0.934	0.814	0.797
Precision	0.664	1.000	0.743	0.000	0.937	0.565	0.662
F-Measure	0.720	1.000	0.709	0.000	0.936	0.667	0.723
Averages							
Recall	0.795	1.000	0.807	0.538	0.781	0.831	0.821
Precision	0.734	1.000	0.856	0.586	0.787	0.757	0.777
F-Measure	0.761	1.000	0.831	0.552	0.784	0.785	0.797

Subsequent to the formal test runs, we perfomed ablation tests to determine the relative impacts on performance by the search, rule, and knowledge-based components. Table 4 reports the results of these tests. The baseline column reports full system average performance for a slightly improved version of the system that participated in the formal tests. The improvements consisted of correcting some programming bugs and making some minor modifications to some of the clustering rules. As the table shows, the baseline results are virtually identical to the average scores in Table 1.

Table 4. Ablation test performance results.

	Baseline	No Knowledge	No Rules/No Knowledge
I2b2 Data			
Recall	0.800	0.803	0.744
Precision	0.878	0.872	0.816
F-Measure	0.834	0.834	0.775
ODIE Data			
Recall	0.841	0.833	0.800
Precision	0.816	0.803	0.749
F-Measure	0.828	0.817	0.772

The "No Knowledge" column reports average performance of the baseline system with all search and rule-based algorithms working, but without the benefit of NLM lexical variant generation and also without the benefit of WordNet's synonyms and hypernyms. Both of these knowledge sources were used in the determination of lexical similarity. Although performance dropped off by approximately 1% for ODIE data, the results were essentially unchanged for the i2b2 data. We conclude from this that the search component is responsible for almost all of the correct lexical similarity determinations.

The "No Rules/No Knowledge" column reports average performance when the rule-based component was disabled. Since the external knowledge sources (NLM, WordNet) were invoked from within the rules, they were also

necessarily disabled. For this test, all candidate similarity matches reported by the search component were accepted without examining them further. As the table shows, performance scores fell by 5.9% for i2b2 data and by 5.6% for ODIE data. This suggests that the rule-based component was responsible for improving overall performance. These results also suggest that the performance of the search component alone, at approximately 77%, affords a good baseline for further improvement by downstream components employing different technologies.

Discussion

The relatively high performance (F-measure=77%) achieved by the "No Rules/No Knowledge" ablated system run demonstrates that lexical similarity is a major component of coreference resolution in the clinical context, and using a search-based component to identify lexically similar candidates is an effective means for identifying initial coreference entities.

Moreover, the 5.6% to 5.9% performance improvement observed when coupling the search component with a rule-based component that both filters candidate coreferential links and also merges coreferential clusters, demonstrates that performance improvements can be made by a lightweight rule-based processing engine operating upon a solid initial foundation furnished by the search component.

The role of external knowledge sources, such as NLM's Lexical Variant Generator and WordNet's synsets and hypenyms, is more difficult to assess. Initially, during system design, we believed that such sources were necessary in order to determine lexical similarity. However, the "No Knowledge" ablation test results suggest that this is not the case. Nevertheless, investigation of the low MUC scores for ODIE "laboratoryortestresult" markables suggests that we do indeed need domain knowledge, and hence external knowledge sources of some kind, to filter out mention associations that would otherwise be made at the lexical level. Accordingly, it would seem fruitful to pursue further performance improvements along these lines.

Conclusion

In this paper we present a search-based approach to clinical text coreference solution that achieves an overall F-measure of 83.3% against i2b2/VA clinical data and 82.7% against ODIE clinical data, with recall and precision values at or above 80% in both cases. The system was configured identically for both annotation schemes and all data sources. Test results demonstrate robustness against the different annotation and tokenization schemes, and also against the different document types. With a search component that by itself achieves an a baseline 77% F-measure, plus a demonstrated 5.6%-5.9% improvement with an active lightweight rule-based processing component and identified areas for improvement, we feel that further investigation along these lines is warranted.

References

1. https://www.i2b2.org/NLP/Coreference/Call.php (accessed August 31, 2011).
2. Uzuner O, et al. 2011 i2b2/VA Co-reference annotation guidelines for the clinical domain.
3. Savona G, Chapman W, Zheng J. Anaphoricity annotation guidelines for the clinical domain.
4. Soon W, Ng H, Lim D. A machine learning approach to coreference resolution of noun phrases. Computational Linguistics 2001;27(4):521-544.
5. Yang X, Su J, Zhou G, Tan CL. An NP-cluster based approach to coreference resolution. In Proceedings of ACL 2004; 127-134.
6. Yang X, Su J, Lang J, Tan CL, Liu T, Li S. An entity-mention model for coreference resolution with inductive logic programming. In Proceedings of ACL-HLT 2008;843-851.
7. Haghighi A, Klein D. Unsupervised coreference resolution in a nonparametric Bayesian model. In Proceedings of ACL 2007;848–855
8. Ng V, Unsupervised models for coreference resolution. In Proceedings of EMNLP 2008;640-649.
9. http://lucene.apache.org/java/docs/index.html (Accessed August 31, 2011).
10. http://wiki.apache.org/lucene-java/PoweredBy (Accessed August 31, 2011).
11. http://lexsrv3.nlm.nih.gov/LexSysGroup/Projects/lvg/current/web/index.html (Accessed August 31, 2011).
12. Fellbaum C, Ed. WordNet: An electronic lexical database. The MIT Press 1998.
13. McCandless M, Hatcher E, Gospodnetic O. Lucene In Action, 2d Ed. Manning Publications Co. 2010;100-101.
14. Begga A, Baldwin B. Algorithms for scoring coreference chains. In Proceedings of the LREC 1998 Workshop on Linguistic Coreference 1998;563-566.
15. Vilain M, Burger J, Aberdeen J, Connelly D, Hirschman L. A model-theoretic coreference scoring scheme. In Proceedings of MUC-6 1995;45-52.
16. Luo X. On coreference resolution performance metrics. In Proceedings of HLT-EMNLP 2005;25-3