A Comparative Study of Co-reference Resolution in Clinical Text

David Hinote1, Carlos Ramirez2, Ping Chen1

1University of Houston - Downtown, Houston, TX; 2University of Puerto Rico, Río Piedra, PR

Abstract

The 2011 I2B2 challenge involves co-reference resolution. Concept mentions have been annotated in clinical texts, and the mentions which co-refer in each document are to be linked by co-reference chains. Normally, there are two ways of constructing a system to automatically discover co-referent links. One is to manually build rules for co-reference (a rule based approach), and the other is to use machine learning algorithms to build the rules automatically (a machine learning approach). There have been systems developed for co-reference resolution by various organizations. The aim of this study was to use the systems which are publicly available, as well as build a system tailored for this challenge using a rule based algorithm, and test these systems on the data provided for this challenge. The study shows the publically available systems do manage to find some of the co-referent links, and the rule based system developed for this challenge performs well finding the majority of the co-referent links. The system that was used to provide the final outputs for the challenge had at highest a 91% overall performance average when cross-checked with the data used for development.

Introduction

Co-reference resolution is the process of linking together concepts which refer to the same entity. In this challenge, hosted by i2b2 (Informatics for Integrating Biology & the Bedside), annotated data has been provided by four institutions: Partners HealthCare, Beth Israel Deaconess Medical Center, The University of Pittsburgh, and the Mayo Clinic. This data includes the original texts for each document, a concepts file for each document which describes each concept mention in the texts, and chain files which identify manually created chains in each of the texts as an example of how chains are to look after processing. The annotated concept mentions to be linked are nouns or descriptive phrases in the medical texts which represent people, actions, objects, or ideas and have been given types accordingly. Examples of the types are problems, people, tests, treatments etc. Each type of concept mention will only co-refer with a concept mention of the same type, with the exception of pronouns which can co-refer with any type of mention1. For this challenge, a study was conducted which examined the performance of three publicly available co-reference systems, as well as a rule based chain building algorithm constructed for this challenge.

**Evaluation Methods**

Each system was evaluated in two ways. The first method was to compare each link with the provided co-reference chain annotations, and count it as correct only if it matches exactly with the provided annotation. With this method, single unlinked concept mentions are not considered, and links that fall in the same chain but skip an antecedent are considered incorrect. Results for individual concept mention types using this method are listed in this paper below each system description. The second method of evaluation is with a script provided by I2B2, which conducts 4 types of examinations of the chain output for each system: B-Cubed5, MUC6, Blanc7, and CEAF8. Overall performance results using this method, again, are listed in this paper after each system description.

Systems used in the study

There are a number of systems publicly available for co-reference resolution that can be found by conducting internet searches on most popular search engines. These systems will discover co-referent links and chains from raw text input. To do this the software uses internal functions to find concepts, and then link them2, 3, 4. The three systems used for testing were chosen for their level of development and usability. In this study BART2, the Stanford co-reference system3, and LingPipe4 were tested on the provided training data.

**BART**

BART is an acronym for Beautiful Anaphora Resolution Toolkit and it was developed from a project done at the 2007 Johns Hopkins Summer workshop2. It is available on the website created for the project: <http://www.bart-coref.org/>. Once set up, text is sent to it through a web service, and output is returned in XML format. The output contains detected concept mentions and if they belong to a chain, the chain identifier is included in the XML tag of the concept mention. The execution time for co-reference resolution using the BART on the clinical documents ranges from 20 seconds to 10 minutes, and depends on the length of the document and the size of the sentences it detects. A translator was created to compare the BART output to the chain files included with the input texts. Only concept mentions detected by the BART system and listed by the I2B2 annotations were considered for testing. All other mentions and co-referent links were discarded. After running the BART system on files from each of the data sets provided, the following performance results were obtained (Table 1).

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Data Set | Overall | People | Problems | Tests | Treatments | All Others |
| Beth Israel | 77.5% | 59% | 20.2% | 16.67% | 30% | N/A |
| Partners Healthcare | 71.2% | 45.7% | 20.6% | 25.3% | 26.3% | N/A |
| Mayo Clinic | 43.5% | 4.1% | 0% | N/A | 0% | 0% |

Table 1. BART performance results. The overall column is the unweighted average of the four metrics provided by I2B2. The other columns are the f1 performance score of the individual concept mention categories evaluated using the first method described above in “Evaluation Methods”.

Stanford Co-reference System

The Stanford co-reference system is an ongoing project by the Stanford Natural Processing Language Group4. It can be found at the Stanford Natural Language Processing Group’s website: <http://nlp.stanford.edu/software/dcoref.shtml>. It uses what is called a “Multi-pass sieve” to perform co-reference resolution, which is a layered approach to detecting links between mentions. It starts with the strongest match first then uses more and more relaxed criteria for matches as it runs down the layers of co-referring rules3. Like BART, it uses its own internal functions to identify concept mentions. Execution time of the Stanford system ranged from 2.5 minutes to 20 minutes per document, and depended on the length of the text. Input and output for this system was done by calling the Java classes directly from the computational program developed for this study. Input was supplying the raw text in a string, and output from this system comes in the form of a map stored in an array. Each element of the array holds the location, in the form of line number and word number in the text, of a source mention, and a destination mention. A simple mapping function was constructed to convert the Stanford concept locations to I2B2 concept locations. Only concept mentions that were found by the Stanford system and listed by the I2B2 annotations were considered, all other mentions and co-referent links were discarded. After running the Stanford system on files from each of the data sets provided, the following performance results were obtained (Table 2).

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Data Set | Overall | People | Problems | Tests | Treatments | All Others |
| Beth Israel | 62.7% | 24.4% | 1.5% | 2.9% | 9.2% | N/A |
| Partners Healthcare | 63.3% | 20.7% | 3.4% | 5.3% | 8.4% | N/A |
| Mayo Clinic | 43.6% | 0.69% | 0% | N/A | 0% | 0% |

Table 2. Stanford performance results. The overall column is the unweighted average of the four metrics provided by I2B2. The other columns are the f1 performance score of the individual concept mention categories evaluated using the first method described above in “Evaluation Methods”.

LingPipe

LingPipe is a suite of natural language processing tools provided by the Alias-i company as a commercial NLP product. It is available at no cost for research purposes at the Alias-i website: <http://alias-i.com/lingpipe>. LingPipe performs Co-reference resolution through a set of heuristic algorithms which link together mentions found by internal functions4. Execution time for LingPipe’s co-reference function per document is between 1 and 5 seconds, making it the fastest of the three systems tested. Input for the system was through command line functions specifying the location of the input text documents, and output was a text document containing xml tags surrounding discovered concept mentions and a chain identifier if the mention was found to be co-referent. A translator similar to the one used to map the BART system output was constructed to make the data useable in this study. After filtering out concept mentions not annotated in the I2B2 data, the following performance results were obtained (Table 3).

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Data Set | Overall | People | Problems | Tests | Treatments | All Others |
| Beth Israel | 62.7% | 24.4% | 1.5% | 2.9% | 9.2% | N/A |
| Partners Healthcare | 63.3% | 20.7% | 3.4% | 5.3% | 8.4% | N/A |
| Mayo Clinic | 42.3% | 0.71% | 0% | N/A | 0% | 0% |

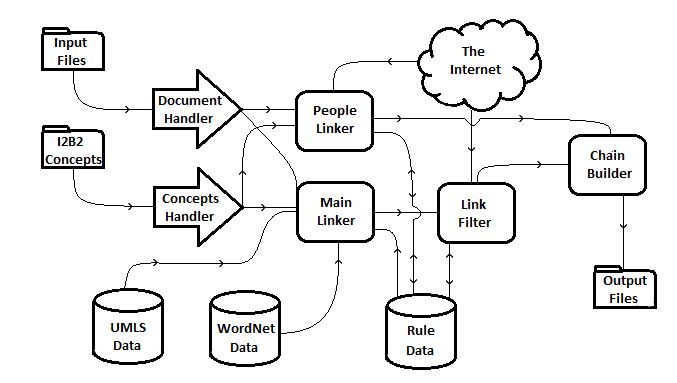
Table 3. LingPipe performance results. The overall column is the unweighted average of the four metrics provided by I2B2. The other columns are the f1 performance score of the individual concept mention categories evaluated using the first method described above in “Evaluation Methods”.

Rule Based Co-reference Algorithm

Because the specifications for co-reference for the i2b2 challenge were well defined, and the type of data provided is specific kind of document1, the type of algorithm to be built for this study is rule based. This algorithm was developed as an additional system to test on the I2B2 data, and to test various ideas for accomplishing the task given by the 2011 I2B2 challenge. The algorithm was developed by examining a sample of files from the pool of data and writing linking functions, or rules, based on observation. The linking functions were checked across the entire data set to get an idea of which rules worked, and which did not. In addition to knowledge-poor string matching rules, the final algorithm uses Wordnet9, the UMLS database10, and automatic internet searches to help classify concept mentions in order to link them.

Concept Handling

All concept types are processed though the same pipeline in the algorithm except for the mentions that have to do with people. These “people” mentions which are not pronouns are categorized as being either the subject of the document, or a third party. In order to make this determination, it was found that internet searches could be used to determine if a mention is of medical personnel or not. That along with coded rules for identifying other third party mentions, such as family, and connecting pronouns to the appropriate names comprises the section of the algorithm that links the mentions having to do with people. All other concept mentions were examined and linked based on semantic similarity. This similarity is determined using several methods. The first method is using string matching functions to match mentions spelled exactly or almost the same and examine phrases between mentions of the same types which would indicate similarity. Another method uses Wordnet synonyms to match words within the mentions. For medical terminology, the UMLS database is used to determine closely related medical mentions. After the semantic links are made, they are passed over to filters to eliminate links that actually refer to two different things based on clues found in the sentences surrounding the mentions in question. These clues can include things such as dates, locations, or descriptive modifiers not included in the span of the mention. These clues are compared through string matching as well as automatic internet searches. The filter portion of the algorithm also eliminates links using Wordnet to determine what kind of word or phrase the mention is, such as removing all adjectives from chains. The entire flow of data for the algorithm can be seen in Figure 1.

****

**Figure 1.** Data flow chart for the rule based algorithm

**Rule Based Algorithm Performance**

The final form of the rule based system developed for this challenge was cross-checked with the I2B2 files provided and had the following performance results (Table 4). Execution time per document ranges from 1 to 10 seconds, and depends on the length of the document.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Data Set | Overall | People | Problems | Tests | Treatments | All Others |
| Beth Israel | 89.1% | 95.8% | 69% | 38.9% | 59.7% | N/A |
| Partners Healthcare | 91.2% | 95.3% | 69.6% | 46.24% | 62.4% | N/A |
| Mayo Clinic | 78.9% | 59.3% | 66.7% | N/A | 50% | 45.3% |

Table 4. Rule based algorithm performance results. The overall column is the unweighted average of the four metrics provided by I2B2. The other columns are the f1 performance score of the individual concept mention categories evaluated using the first method described above in “Evaluation Methods”.

**Combining Results**

Once result data was collected, combinations of link results from the rule based system and the BART system were examined since the BART system showed the highest amount of correct link predictions. After combining the results from the two systems as a union of the sets, the statistics showed an increase of about 1% in recall but a decline of about 15% in precision, bringing the f1 score down overall.

**Conclusion**

Since the goal of the 2011 I2B2 shared task was to mark concept mentions as co-referent or not, the rule based system developed for this study was used to mark links in the test data released by the organization for the challenge. This decision was made based on the results from cross-checking the performance of each system on the training data provided. The results show the BART system performed the best out of the three publicly available co-reference systems tested in this study on this specific collection of data. The results also show that manually creating rules for co-reference based on observation of training data is a valid way to accomplish this co-reference task, and in this case performed well using the guidelines laid out by the hosts of the competition.

**Acknowledgements**

This work is funded by National Science Foundation grant CNS 0851984 and Department of Homeland Security grant 2009-ST-061-C10001.

References

1. Uzuner, O, et al. 2011 i2b2/VA co-reference annotation guidelines for the clinical domain. <https://www.i2b2.org/NLP/Coreference/assets/>. (2011).
2. Poesio, M., Ponzetto, S., Versley, Y., Eidelman, V., Jem, A., Moschitti, A., Yang, X., Rodriguez, K., Uryupina, O. BART coreference system. <http://www.bart-coref.org/>. (2007).
3. Karthik Raghunathan, Heeyoung Lee, Sudarshan Rangarajan, Nathanael Chambers, Mihai Surdeanu, Dan Jurafsky, Christopher Manning. A multi-pass sieve for coreference resolution. EMNLP-2010, Boston, USA. (2010).
4. Alias-i. LingPipe 4.1.0. <http://alias-i.com/lingpipe>. (2008).
5. Bagga, Amit. Baldwin, Breck. Algorithms for scoring coreference chains. In Proceedings of the Linguistic Coreference Workshop at The First International Conference on Language Resources and Evaluation (LREC'98). (1998).
6. Vilain, Marc, et al. A model-theoretic coreference scoring scheme. Proceedings of the Sixth Message Understanding Conference (MUC-6). (1995).
7. Lita, Lucian Vlad. Rogati, Monica. Lavie, Alon. Blanc: Learning evaluation metrics for mt. In

Proceedings of Human Language Technology Conference and Conference on Empirical Methods in

Natural Language Processing. (2005).

1. Luo, X. On coreference resolution performance metrics. In HTL-EMNLP. (2005).
2. Fellbaum, Christiane. WordNet: An Electronic Lexical Database. Cambridge, MA: MIT Press. (1998)
3. U.S. National Library of Medicine. Unified Medical Language System. <http://www.nlm.nih.gov/research/umls/>. (2011).