Domain Adaptation of Coreference Resolution for Radiology Reports

Emilia Apostolova, Noriko Tomuro, Pattanasak Mongkolwat*, Dina Demner-Fushman[∓]
College of Computing and Digital Media, DePaul University, Chicago, IL
*Department of Radiology, Northwestern University Medical School, Chicago, IL
[∓]Communications Engineering Branch, National Library of Medicine, Bethesda, MD
emilia.aposto@gmail.com, tomuro@cs.depaul.edu,
p-mongkolwat@northwestern.edu, ddemner@mail.nih.gov

Abstract

In this paper we explore the applicability of existing coreference resolution systems to a biomedical genre: radiology reports. Analysis revealed that, due to the idiosyncrasies of the domain, both the formulation of the problem of coreference resolution and its solution need significant domain adaptation work. We reformulated the task and developed an unsupervised algorithm based on heuristics for coreference resolution in radiology reports. The algorithm is shown to perform well on a test dataset of 150 manually annotated radiology reports.

1 Introduction

Coreference resolution is the process of determining whether two expressions in natural language refer to the same entity in the world. General purpose coreference resolution systems typically cluster all mentions (usually noun phrases) in a document into coreference chains according to the underlying reference entity. A number of coreference resolution algorithms have been developed for general texts. To name a few, Soon et al. (2001) employed machine learning on the task and achieved an F-score of 62.6 and 60.4 on the MUC-6 (1995) and MUC-7 (1997) coreference corpora respectively. Ng et al. (2002) improved this learning framework and achieved F-scores of 70.4 and 63.4 respectively on the same datasets.

There are also a number of freely available off-the-shelf coreference resolution modules developed for the general domain. For example, BART (Versley et al., 2008) is an open source coreference resolution system which provides an implementation of the Soon et al. algorithm (2001). The Stanford Deterministic Coreference Resolution System (Raghunathan et al., 2010) uses an unsupervised sieve-like approach to coreference resolution. Similarly, the GATE Information Extraction system (Cunningham et al., 2002) includes a rule-based coreference resolution module consisting of orthography-based patterns and a pronominal coreferencer (matching pronouns to the most recent referent).

While coreference resolution is a universal discourse problem, both the scope of the problem and its solution could vary significantly across domains and text genres. Newswire coreference resolution corpora (such as the MUC corpus) and general purpose tools do not always fit the needs of specific domains such as the biomedical domain well.

The importance and distinctive characteristics of coreference resolution for biomedical articles has been recognized, for example (Castano et al., 2002; Gasperin, 2006; Gasperin et al., 2007; Su et al., 2008). Within the biomedical field, clinical texts have been noted as a genre that needs specialized coreference corpora and methodologies (Zheng et al., 2011). The importance of the task for the clinical domain has been attested by the 2011 i2b2 NLP shared task (Informatics for Integrating Biology and the Bedside[1]) which provided an evaluation platform for coreference resolution for clinical texts.

However, even within the clinical domain, coreference in different sub-genres could vary signifi-

[1] https://www.i2b2.org/NLP/Coreference/

Proceedings of the 2012 Workshop on Biomedical Natural Language Processing (BioNLP 2012), pages 118–121,
Montréal, Canada, June 8, 2012.

cantly. In this paper we demonstrate the idiosyncrasies of the task of coreference resolution in a clinical domain sub-genre, radiology reports, and describe an unsupervised system developed for the task.

2 Coreference Resolution for Radiology Reports

Radiology reports have some unique characteristics that preclude the use of coreference resolution modules or algorithms developed for the general biomedical domain or even for other types of clinical texts. The radiology report is a clinical text used to communicate medical image findings and observations to referring physicians. Typically, radiology reports are produced by radiologists after examining medical images and are used to describe the findings and observations present in the accompanied images.

The radiology report accompanies an imaging study and frequently refers to artifacts present in the image. In radiology reports, artifacts present in the image exhibit *discourse salience*, and as a result are often introduced with definite pronouns and articles. For example, consider the sentence *The pericardial space is clear*. The definite noun phrase *the pericardial space* does not represent an anaphoric (or cataphoric) discourse entity and has no antecedent. In contrast, coreference resolution in general texts typically considers definite noun phrases to be anaphoric discourse entities and attempts to find their antecedents.

Another important distinction between general purpose coreference resolution and the coreference resolution module needed by an NLP system for clinical texts is the scope of the task. General purpose coreference resolution systems typically cluster all mentions in a document into coreference chains. Such comprehensive mention clustering is often not necessary for the purposes of clinical text NLP systems. Biomedical Information Extraction systems typically first identify named entities (medical concepts) and map them to unambiguous biomedical standard vocabularies (e.g. UMLS[2] or RadLex[3] in the radiological domain). While multiple mentions of the same named entity could exist in a document, in most cases these mentions were previously assigned to the same medical concept. For example, multiple report mentions of '*the heart*' or '*the lung*' will normally be mapped to the same medical concept and clustering of these mentions into coreference chains is typically not needed.

3 Task Definition

Analysis revealed that the coreference resolution task could be simplified and still meet the needs of most Information Extraction tasks relevant to the radiological domain. Due to their nature, texts describing medical image finding and observations do not contain most pronominal references typically targeted by coreference resolution systems. For example, no occurrence of personal pronouns (e.g. *he, I*), possessive pronouns (e.g. *his, my*), and indefinite pronouns (e.g. *anyone, nobody*) was found in the validation dataset. Demonstrative pronouns and non-pleonastic 'it' mentions were the only pronominal references observed in the dataset[4]. The following examples demonstrate the use of demonstrative pronouns and the non-pleonastic 'it' pronoun (shown in bold):

There is prominent soft tissue swelling involving the premaxillary tissues. ***This*** *measures approximately 15 mm in thickness and extends to the inferior aspect of the nose.*

There is a foreign object in the proximal left mainstem bronchus on series 11 image 17 that was not present on the prior study. ***It*** *has a somewhat ovoid to linear configuration.*

Following these observations, the coreference resolution task has been simplified as follows. Coreference chains are assigned only for demonstrative pronouns and 'it' noun phrases. The coreference resolution task then involves selecting for each mention a single best antecedent among previously annotated named entities (medical concepts) or the NULL antecedent.

4 Dataset

A total of 300 radiology reports were set aside for validation and testing purposes. The dataset consists

[2]http://www.nlm.nih.gov/research/umls/

[3]http://www.radlex.org/

[4]Pleonastic 'it' refers to its use as a 'dummy' pronoun, e.g. *It is raining*, while non-pleonastic use of the pronoun refers to a specific entity.

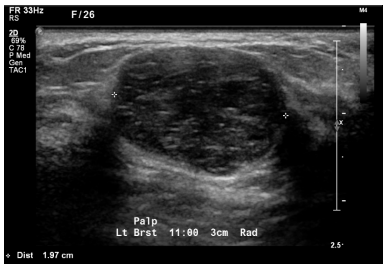

Figure 1: A sample DICOM image from an imaging study described by the following radiology report snippet: . . . *FINDINGS: Targeted sonography of the upper inner left breast was performed. At the site of palpable abnormality, at the 11 o'clock position 3 cm from the nipple, there is an oval circumscribed, benign-appearing hypoechoic mass measuring 2.0 x 1.6 x 1.4 cm. There is mild internal blood flow. It is surrounded by normal appearing glandular breast tissue.*. . .

of 100 Computed Tomography Chest reports, 100 Ultrasound Breast reports, and 100 Magnetic Resonance Brain reports, all randomly selected based on their report types from a dataset of more than 100,000 de-identified reports spanning a period of 9 years[5]. These three types of reports represent a diverse dataset covering representative imaging modalities and body regions. Figure 1 shows a sample Breast Ultrasound DICOM[6] image and its associated radiology report.

The reports were previously tagged (using an automated system) with medical concepts and their semantic types (e.g. anatomical entity, disorder, imaging observation, etc.). Half of the dataset (150 reports) was manually annotated with coreference chains using the simplified task definition described above. The other half of the dataset was used for validation of the system described next.

5 Method and Results

The coreference resolution task involves selecting for each mention a single best antecedent among previously annotated named entities or the NULL antecedent. Mentions are demonstrative pronoun phrases or definite noun phrases containing previously annotated named entities.

We implemented an algorithm for the task described above which was inspired by the work of Haghighi and Klein (2009). The algorithm first identifies mentions within each report and orders them linearly according to the position of the mention head. Then it selects the antecedent (or the NULL antecedent) for each mention as follows:

1. The possible antecedent candidates are first filtered based on a distance constraint. Only mentions of interest belonging to the preceding two sentences are considered. The rationale for this filtering step is that radiology reports are typically very concise and less cohesive than general texts. Paragraphs often describe multiple observations and anatomical entities sequentially and rarely refer to mentions more distant than the preceding two sentences.

2. The remaining antecedent candidates are then filtered based on a syntactic constraint: the coreferent mentions must agree in number (singular or plural based on the noun phrase head).

3. The remaining antecedent candidates are then filtered based on a semantic constraint. If the two mentions refer to named entities, the named entities need to have the same semantic category[7].

4. After filtering, the closest mention from the set of remaining possible antecedents is selected. If the set is empty, the NULL antecedent is selected.

Pairwise coreference decisions are considered transitive and antecedent matches are propagated transitively to all paired co-referents.

The algorithm was evaluated on the manually annotated test dataset. Results (Table 1) were computed using the pairwise F1-score measure: precision, recall, and F1-score were computed over all pairs of mentions in the same coreference cluster.

Precision	Recall	F1-score
74.90	48.22	**58.66**

Table 1: Pairwise coreference resolution results.

The system performance is within the range of state-of-the-art supervised and unsupervised coreference resolution systems[8]. F1-scores could range

[5]The collection is a proprietary dataset belonging to Northwestern University Medical School.

[6]Digital Imaging and Communications in Medicine, © The National Electrical Manufacturers Association.

[7]The same semantic type in the case of UMLS concepts or the same parent in the case of RadLex concepts.

[8]Source code for the described system will be made available upon request.

between 39.8 and 67.3 for various methods and test sets (Haghighi and Klein, 2009). The simplification of the coreference resolution problem described above allowed us to focus only on coreference chains of interest to clinical text Information Extraction tasks and positively influenced the outcome. In addition, our goal was to focus on high precision results as opposed to optimizing the overall F1-score. This guarantees that coreference resolution errors will result in mostly omissions of coreference pairs and will not introduce information extraction inaccuracies.

6 Conclusion

In this paper, we presented some of the challenges involved in the task of adapting coreference resolution for the domain of clinical radiology. We presented a domain-specific definition of the coreference resolution task. The task was reformulated and simplified in a practical manner that ensures that the needs of biomedical information extraction systems are still met. We developed an unsupervised approach to the task of coreference resolution of radiology reports and demonstrate state-of-the-art precision and reasonable recall results. The developed system is made publicly available to the NLP research community.

References

J. Castano, J. Zhang, and J. Pustejovsky. 2002. Anaphora resolution in biomedical literature. In *International Symposium on Reference Resolution*. Citeseer.

D.H. Cunningham, D.D. Maynard, D.K. Bontcheva, and M.V. Tablan. 2002. GATE: A Framework and Graphical Development Environment for Robust NLP Tools and Applications.

C. Gasperin, N. Karamanis, and R. Seal. 2007. Annotation of anaphoric relations in biomedical full-text articles using a domain-relevant scheme. In *Proceedings of DAARC*, volume 2007. Citeseer.

C. Gasperin. 2006. Semi-supervised anaphora resolution in biomedical texts. In *Proceedings of the Workshop on Linking Natural Language Processing and Biology: Towards Deeper Biological Literature Analysis*, pages 96–103. Association for Computational Linguistics.

A. Haghighi and D. Klein. 2009. Simple coreference resolution with rich syntactic and semantic features. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 3-Volume 3*, pages 1152–1161. Association for Computational Linguistics.

V. Ng and C. Cardie. 2002. Improving machine learning approaches to coreference resolution. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 104–111.

K. Raghunathan, H. Lee, S. Rangarajan, N. Chambers, M. Surdeanu, D. Jurafsky, and C. Manning. 2010. A multi-pass sieve for coreference resolution. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 492–501. Association for Computational Linguistics.

W.M. Soon, H.T. Ng, and D.C.Y. Lim. 2001. A machine learning approach to coreference resolution of noun phrases. *Computational Linguistics*, 27(4):521–544.

J. Su, X. Yang, H. Hong, Y. Tateisi, J. Tsujii, M. Ashburner, U. Leser, and D. Rebholz-Schuhmann. 2008. Coreference resolution in biomedical texts: a machine learning approach. *Ontologies and Text Mining for Life Sciences 08*.

Y. Versley, S.P. Ponzetto, M. Poesio, V. Eidelman, A. Jern, J. Smith, X. Yang, and A. Moschitti. 2008. Bart: A modular toolkit for coreference resolution. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies: Demo Session*, pages 9–12. Association for Computational Linguistics.

J. Zheng, W.W. Chapman, R.S. Crowley, and G.K. Savova. 2011. Coreference resolution: A review of general methodologies and applications in the clinical domain. *Journal of biomedical informatics*.