Acquiring Domain Specific Knowledge and Coreference Cues for Coreference Resolution

Nathan Gilbert
School of Computing
University of Utah
Salt Lake City, UT 84112
ngilbert@cs.utah.edu

Introduction

Coreference Resolution is the task of determining which noun phrases in a text refer to the same entity. In Figure 1, all coreferent entities are labeled with the same subscript value. The noun phrases { *Minister Monica de Greiff*, *her*, *she* } all refer to the same individual, therefore they are all coreferent and are said to be part of a *coreference chain*.

> *Minister Monica de Greiff*$_1$ briefed *President Barco*$_2$ today on *her*$_1$ visit to the United States. *The president*$_2$ was pleased with the report *she*$_1$ presented.

Figure 1: Coreference Chain Example

The goal of a coreference resolution (CR) system is take unstructured text and produce coreference chains of all the entities within it. This task is normally broken down into subordinate steps, such as detecting what entities can participate in coreference relations and what links between entities can be determined.

This research attempts to address several problems found in general coreference resolution systems and presents new methods for making better resolutions in a domain specific setting.

Motivation & Goals

Current approaches of automatic CR generally involve some manner of machine learning algorithm utilizing a set of general, knowledge-poor features. These feature sets rely heavily on noun phrase recency, semantic class information, and string match characteristics as in (Ng and Cardie 2002). Many current systems make resolution decisions on a pair-by-pair basis between noun phrases in a document, independent of any previous decisions, with a clustering step at the end to create coreference chains. Clustering approaches do exist (Cardie and Wagstaff 1999), and they attempt to build coreference chains simultaneously, but utilize the same general, knowledge poor feature sets as the pairwise approaches.

In our paper (Stoyanov et al. 2009), we found that the difficulty that current CR systems have with a particular corpus is related to the number of resolutions involving common nouns such as "company" or "man". Common nouns are often associated with concepts or used in place of entity names, both of which are arguably domain dependent.

Knowing what to do with common nouns in a domain is tricky. There are instances where many common nouns may appear in a document, yet are not coreferent, such as sports articles which reference many different "games". What we show is that learning how common nouns are involved in coreference relations would go a long way in pushing the current state of coreference research beyond general and knowledge poor approaches.

There are specific sets of entities that need domain specific knowledge. General semantic resources often will not contain the domain specific knowledge required to make some resolutions. "Ebola Hemorrhagic Fever" is very likely to be coreferent with the phrase "the deadly virus" in a disease outbreak domain, whereas most general purpose resolvers are unlikely to have any relevant semantic information for either noun phrase.

Knowledge from the "real world" is often needed as well, and can be viewed as a type of domain specific information. For instance, in Figure 2, it is very unlikely that any general semantic dictionary would know that "Mario Vargas Llosa" is a novelist. The relationship between "novelist" and "Mario Vargas Llosa" is known outside of a Latin-American terrorist domain, but not every occurrence of "novelist" in every domain is going to be coreferent to Mr. Llosa. The knowledge that Mr. Llosa is a novelist mentioned frequently *in this domain* should be useful.

> The Peruvian navy formally denied that it had uncovered a plot to assassinate *Mario Vargas Llosa*$_1$. . . . The report states that a MRTA hit squad would murder *the novelist*$_1$ . . .

Figure 2: Resolution requiring domain knowledge

Learning what noun phrases are likely to corefer would be of great benefit to training models over a new domain for which semantic information is missing or annotated training data is limited. Confining ourselves to specific domains may direct our search towards higher precision, and acquiring more relevant information between coreferent noun phrases.

Proposed Work

The first phase of this work will be to explore the power of domain specific features in CR systems. I have found that by using models trained on very general corpora to classify on more domain specific document sets, the results are surprising. It is the case that the models trained on generic texts perform better than the models trained on the texts from the very domain which is being tested. My hypothesis is that the "same-domain" models are less effective on data requiring strong domain information because there is no domain specific knowledge present in their models.

The solution to this problem is to introduce features that incorporate domain specific knowledge. Learning what noun phrases commonly corefer in a given domain is one way to do this. It makes sense to learn that "the president" may often refer to "Barack Obama" in a given collection of texts. Some approaches to this have been tried in (Rahman and Ng 2011), but surprisingly did little to help CR performance. The authors used general, non-domain specific corpora used in these experiments. My latest results have shown the true power of domain specific features is realized when they are applied to strongly domain related texts.

Another approach is to utilize resources such as the Sundance shallow parser (Riloff and Phillips 2004) to learn what semantic classes specific nouns (especially common nouns) refer to in a given domain. Learning that "the case" often refers to a person in a biomedical domain could help make resolutions by improving semantic class agreement.

Contextual cues shared between coreferent entities could be another approach to utilize lexicalization for CR. For instance, in a sports domain, learning that th phrase: "the game" is often *not* coreferent with other instances of "the game", but are instead coreferent with noun phrases found in a particular context such as "*last Tuesday's* game".

Two major problems confront the previous approaches, and they will compose phase two. One problem is the "multi-match" scenario. This occurs when there is more than a single possible lexical or semantic match present as a possible antecedent. An example is a document that contains both the noun phrases "El Salvador" and "Panama" in which the noun phrase "the country" could possibly be linked to. From a domain knowledge standpoint, both antecedents are possible, but obviously, at least one is incorrect.

Solutions to the "multi-match" problem may require more attention to be paid to the discourse properties of a document. Utilizing discourse parsers, or creating a focus model of the entities participating in a document may overcome this challenge. Another solution may lie in designing a different approach to the classification of resolutions themselves. Instead of looking at each classification in isolation, consider clustering entities from the start but by utilizing new domain specific features.

The second problem is how to deal with sparse data. All the approaches discussed above require some annotated data. For any domain, the number of entities we've witnessed as coreferent in the annotated data set is going to be small, we need a way of acquire more without annotating more documents.

One way of getting a large amount of unannotated resolutions via heuristic approaches such as those from (Bean and Riloff 2004; Baldwin 1997). These methods are very precise, but suffer from low recall. They also find more general case resolutions and fewer of the domain-dependent resolutions involving common nouns that we need. New heuristics based on discourse focus or utilizing external knowledge-rich sources such as the web or other domain specific resources may become important in this phase. Increasing the number of positive (or negative) links gathered from domain specific unannotated documents is a necessary step of this research.

Current Progress

I am roughly 75% complete with the first phase of this work as of February 2012. I have incorporated domain knowledge that improves performance over previous attempts. I have yet to look in earnest at utilizing contextual clues which is the last step of phase one. All of this work has been done by myself with guidance from my advisor, Ellen Riloff.

I hope to have completed the first phase and have a substantial start on the second phase by July. My plan to complete and defend this work by December 2013.

In summary, we have explored current CR systems and have identified the lack of domain specific knowledge as a major problem. I have has successfully acquired domain specific knowledge from annotated documents that supports this claim. To continue this research, two subproblems of this domain specific approach have been identified and paths for attacking them have been outlined.

References

Baldwin, B. 1997. Cogniac : High precision coreference with limited knowledge and linguistic resources. *Proceedings of a Workshop on Operational Factors in Practical, Robust Anaphora Resolution for Unrestricted Texts*.

Bean, D., and Riloff, E. 2004. Unsupervised learning of contextual role knowledge for coreference resolution. *Proceedings of HLT/NAAL 2004*.

Cardie, C., and Wagstaff, K. 1999. Noun phrase coreference as clustering. *Proceedings of the Joint SIGDAT Conference on Empirical Methods in Language*.

Ng, V., and Cardie, C. 2002. Improving machine learning approaches to coreference resolution. *Proceedings of the 40th Annual Meeting of the ACL* 104–111.

Rahman, A., and Ng, V. 2011. Coreference resolution with world knowledge. *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics and Human Language Technologies (ACL-HLT)* 814–824.

Riloff, E., and Phillips, W. 2004. An introduction to the sundance and autoslog systems. Technical report uucs-04-015, School of Computing, University of Utah.

Stoyanov, V.; Gilbert, N.; Cardie, C.; and Riloff, E. 2009. Conundrums in noun phrase coreference resolution: Making sense of the state-of-the-art. *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th IJCNLP (ACL-IJCNLP 2009)*.