Exploring Knowledge-Rich Solutions to Noun Phrase Coreference Resolution

Nathan Gilbert & Ellen Riloff
University of Utah

Introduction

Coreference resolution is the task of identifying coreferent expressions in text.

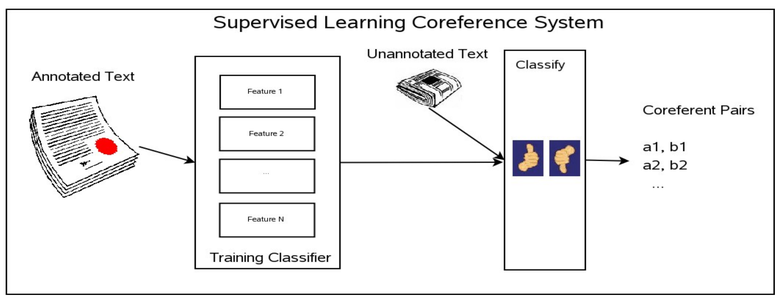
Accurate coreference resolution can improve other tasks such as machine translation, information retrievel and document summarization.

Currently, the best approaches involve some form of supervised Machine Learning algorithms, which requires annotated corpora.

This requirement is expensive and time consuming.

The initial stage of this project was to implement a state of the art supervised learning based coreference system.

The general flow of this system is presented next, notice that for supervised learning systems, the classifier must be trained on annotated text.



This work is still on going at the University of Utah, the following is an overview of previous work and our motivation for this project.

What is coreference?

Two textual entities that refer to the same object in "the real world."

More Formal Definition:

- α_1 and α_2 corefer if and only if Referent(α_1) = Referent(α_2).
- Referent(α) is 'the entity referred to by α.'

Coreference is therefore an equivalence relation for the following three properties hold: reflexive, symmetric and transitive.

Often α_1 and α_2 are referred to as the antecedent and anaphor respectively.

Specific examples of definite noun phrases are always coreferential:

David Beckham is *the L.A. Galaxy midfielder.*

The proper name *David Beckham* and the definite noun phrase *the L.A. Galaxy midfielder* are coreferential by the above definition.

Appositives are also good examples of coreference:

Mary, vice president of public affairs, ...

The current state of the art

A lot of work has gone into building coreference systems that relied on supervised learning to classify textual entities as coreferent.

One such system that has garnered a lot of attention is the work of Soon et al. [1] The following figure displays the system pipeline from free text to "Markables", which was the name they gave to possible coreferential entities.

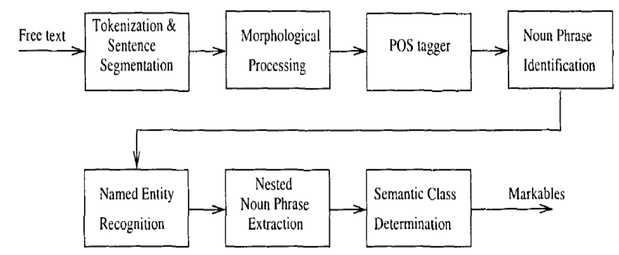


Figure 1
System architecture of natural language processing pipeline.

From here, the markables would be passed to a Decision Tree based classifier that had been previously trained on annotated text.

The results of this system on the Message Understanding Conference 6 & 7 (MUC) data sets are below:

	Recall	Precision	F-Measure
MUC-6	58.6%	67.3%	62.6%
MUC-7	56.1%	65.5%	60.4%

Reconcile

At the University of Utah we have developed our own fully automatic coreference resolution system, named **Reconcile**.

The approach taken in this project is similar to that of Cardie & Ng [2]. We have developed a feature set heavily influence by their work but differ in the modularity and range experiments that are possible with Reconcile.

In short:

- A supervised learning coreference resolution system
- extracts noun phrases on its own
- trains and tests a model based off of these extractions
- various different classification algorithms can be used

Latest Scores:

	Recall	Precision	F-Measure
MUC-6	67.7%	68.0%	67.8%
MUC-7	57.9%	72.6%	64.4%

Reconcile will soon be publicly available.

No Annotated Data?

Annotated data is expensive and combersome to obtain.

Switching Domains can be problematic without new data.

The next steps of this research involve determing methods for solving this problem.

Many subsystems for coreference systems are domain dependent (i.e. they've been trainined on a specific domain.)

This can have a negative impace on a coreference system on a new domain.

Are there unsupervised or semi-supervised methods for attacking this problem?

Coreference and the Web

Improving coreference by adding in "real-world" knowledge.

Current coreference systems primarily consist of what can be considered "knowledge-poor" feature sets. What this means is that majority of characteristics that each classification is based do not require "real-world" information.

Can we leverage information gathered from search engines to help with coreference?

Method: Plug a possible anaphor and antecedent pair into your favorite search engine.

The knowledge coming from the Web has take the form of:

- co-occurrence statistics between anaphor and antecedent
- Point wise mutual information between the anaphor and antecedent counts
- Grep-like patterns for capturing the anaphor and antecedent in snippets return from the search engine:
  - "Bill Gates *is* CEO" - predicate nominal relationship
  - "Bill Gates, Micrsoft CEO" - Appositives relationship
- Co-occurence of patterns taken from the snippets returned by the search engine.

Unfortunately, these methods have not yet shown a significant improvement to coreference resolution.

References

(1) Wee Meng Soon, Hwee Tou Ng, and Daniel Chung Yong Lim. A machine learning approach to coreference resolution of noun phrases. HLT/NAACL, 2004.
(2) Ng, V., and Cardie, C. Improving machine learning approaches to coreference resolution. Proceedings of the 40th Annual Meeting of the ACL, 2002.

Ellen Riloff: riloff@cs.utah.edu
Nathan Gilbert: ngilbert@cs.utah.edu

THE UNIVERSITY OF UTAH

This work is a collaboration between the University of Utah, Cornell University and Lawrence Livermore National Labs.