Cross-Document Coreference Resolution for Entities and Events

Ph.D. Thesis Proposal

Chris Tanner

May 16, 2018

Abstract of “Cross-Document Coreference Resolution for Entities and Events”

Abstract Here

Contents

Chapter 1

Introduction

Thesis Statement: I propose a novel, neural-based mention-pair model for cross-document coreference resolution for events, which uses few lexical features and addresses shortcomings of traditional clustering approaches. I will extend this work by jointly modelling both entities and events, while using structured information (e.g, parse trees). Last, we aim to improve mention detection, whereby we develop an all-inclusive, end-to-end system which jointly resolves mention boundaries and coreference predictions.

1.1 Problem Statement

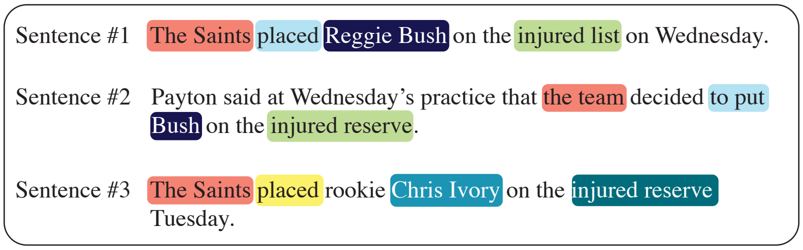
Coreference resolution is the task of identifying – within a single text or across multiple documents – which *mentions* refer to the same underlying discourse object.

A **mention** is a particular instance of word(s) in a document which represent an *entity* or *event*, such as *Barack Obama*, *he*, or *announced*.

An **entity** may be a person, location, time, or an organization. The mentions which refer to them may be *named*, *nominal*, or *pronominal*:

- Named mentions are represented by proper names (e.g., André Benjamin or Pakse, Laos)
- Pronominal mentions are represented by pronouns (e.g., she or it)
- Nominal mentions are represented by descriptive words, not composed entirely of a named entity or pronouns (e.g., The well-spoken citizen)

An **event** can generally be thought of as a specific action. Quine [1] was the first to propose that an event refers to a physical object which is grounded to a specific time and location, and that two events are identical (i.e., co-referent) if they share the same spatiotemporal location. This definition has become the general consensus within the community. Specifically, two co-referent events must share the same *properties* and *participants*. For example, in Figure 1.1, sentences #1 and #2 contain the co-referent events ("placed" and "put"), yet neither are co-referent with events in sentence #3. Often times, the participants (arguments) may be referred to in different ways, implied, or missing altogether.



Sentence #1 The Saints placed Reggie Bush on the injured list on Wednesday.

Sentence #2 Payton said at Wednesday's practice that the team decided to put Bush on the injured reserve.

Sentence #3 The Saints placed rookie Chris Ivory on the injured reserve Tuesday.

Figure 1.1: Sample of a coreference resolution corpus (ECB+), depicting gold coref mentions as having shared box colors.

Coreference resolution is concerned with linking either entities together and/or events together; that is, entities shall not be linked to events, and doing so would be considered an incorrect link. Although one may be interested in evaluating coreference systems by their ability to correctly link *pairs* of mentions [3], coreference resolution is ultimately a clustering task, whereby we wish to group all like-mentions together, as shown with colored boxes in Figure 1.1. Specifically, coreference systems aim to find a globally-optimal fit of mentions to clusters, whereby every mention m in the corpus is assigned to exactly one cluster C, such that every $m_i, m_j \in C$ are co-referent with each other. If a given m_i is not anaphoric with any other m_j, then it should belong to its own C with a membership of one.

Given a corpus of text documents, coreference resolution can be performed and evaluated on either a **within-document** or **cross-document** basis:

- **Within-document** is when each mention may only link to either (1) no other mention; or (2) other mentions which are contained in the same document. Even if the gold truth data denotes a mention should link with a mention from a different document, we ignore these links during the evaluation.

- **Cross-document** is when the entire corpus is available for linking; a mention is eligible to be co-referent with mentions in any other document, and the evaluation reflects the same. As described in [2], cross-document evaluation is normally conducted by transforming the entire corpus into a “meta-document."

1.2 Coreference Systems

1.2.1 Mention Detection

1.2.2 Coreference Resolution

1.3 Motivation

Bibliography

[1] W.V. O. Quine. Events and reification. In *Action and Events: Perspectives on the philosophy of Donald Davidson*, pages 162–171, 1985. Page 3.

[2] Shyam Upadhyay, Nitish Gupta, Christos Christodoulopoulos, and Dan Roth. Revisiting the evaluation for cross document event coreference. In *COLING*, 2016. Page 4.

[3] Travis Wolfe, Mark Dredze, and Benjamin Van Durme. Predicate argument alignment using a global coherence model. In *Human Language Technologies: Conference of the North American Chapter of the Association of Computational Linguistics, Proceedings*, 2015. Page 4.