Entity-Based Noun Phrase Coreference Resolution

Xiaofeng Yang, Jian Su, and Lingpeng Yang

Institute for Infocomm Research, 21 Heng Mui Keng Terrace, Singapore, 119613
{xiaofengy,sujian,lpyang}@i2r.a-star.edu.sg

Abstract. In this paper we propose an NP coreference resolution system which does resolution on the entity-level. The framework of the system is presented and different resolution strategies are investigated.

1 Introduction

Coreference resolution is the process of linking multiple expressions which refer to the same entity. Traditional supervised machine learning approaches (e.g. [1–3]) do resolution based on the mention-level. Specifically, a pairwise classifier is learned and used to determine whether or not two NPs in a document refer to the same entity in the world. However, as an individual mention usually lacks adequate information about its referred entity (e.g, we could not know the gender or the name of "the president"), it is often difficult to determine whether or not two NPs refer to the same entity simply from the pair itself. Recent research ([4, 5]) has revealed that entity information could help resolution. In our work we would like to further study how to effectively incorporate the entity information into coreference resolution. The framework of such a entity-based system is presented and different resolution strategies are investigated in this paper.

2 Baseline: A Mention-based System

We built a Mention-Mention based system as the baseline, which adopts a learning framework similar to the paradigm proposed by Soon et al. [2].

Each instance takes the form of $\boldsymbol{i}\{\mathrm{NP}_i, \mathrm{NP}_j\}$, which is associated with a feature vector consisting of 12 features ($f_1 \sim f_{12}$) as described in Table 1. During training, for each anaphor NP_j in a given text, a positive instance is generated by pairing NP_j with its closest antecedent. A set of negative instances is also formed by NP_j and each NP occurring between NP_j and NP_i.

When the training instances are ready, a classifier is learned by C5.0 algorithm [6]. During resolution, each encountered noun phrase, NP_j, is paired in turn with each preceding noun phrase, NP_i. For each pair, a testing instance is created and then presented to the decision tree, which returns a confidence value (CF) indicating the likelihood that they co-refer. NP_j will be linked to the NP with the maximal CF (above 0.5).