UNIVERSITY OF OSLO
FACULTY OF HUMANITIES

What is co-reference?

What is the co-reference working group trying to achieve?

A short introduction by Øyvind Eide

The co-referende working group

Scholars, researchers, museum curators or library cataloguers continuously trace co-references in their daily work. This extremely time-consuming work can be made more efficient by sharing the knowledge. Such sharing would complement centralised approaches found in present Digital Library Systems.

The long term goal of the Co-reference Working Group is to research into possible implementations of a Co-reference network for cultural and natural heritage information.

Definitions

Co-reference as a concept is closely related to reference. Definition of reference from Oxford English Dictionary:

- 1. d. Logic and Linguistics. The act or state of referring through which one term or concept is related or connected to another or to objects in the world; also as objective reference, and attrib. as reference class, property.

In our context, co-reference occurs when two information objects — in the cases we will consider, strings — are referring to the same object, usually in the real world.

Co-reference is an important issue internally in culture heritage system. But in this WG, we will concentrate on co-reference as it occurs between organisations, in order to connect objects in one database to objects in other databases.

Historical background

Tracking down co-reference has always been one of the practical tasks performed by researchers, conservators, librarians, and others processing information about real world objects described in texts. An authority list is commonly used to co-reference the authors of different books. Footnotes are commonly used to inform the reader that "the person called John here, is the same as the one called Johannes in this other book".

The aim of our work is to find ways to express similar information in a digital age. It should be expressed in a machine readable way so that it can be used for many different purposes apart from being read by a human. It should be followed by information about the event of making this statement: Who is responsible for the statement, be it a person, a group of people or a computer program? On which grounds were it made? (book references, algorithm, etc.)

Implementation

There are many possible ways to implement co-reference systems. Many of the features described here is already in use many places, e.g. in library or museum management systems. This Working Group will mainly work with systems crossing borders: Between different organisations, between different thematic areas, between different nations.

Hierarchical and distributed co-reference

There are two different ways to see co-reference: hierarchical and distributed.

Hierarchical implementation: An example of a hierarchical system would be a system where all museums in a region created a common authority service for persons. Each museum would then link all references to a certain person to this person's record in the authority service, using an URN reference. Then this regional system could be a node in a national museum system, being a node in a national cultural system, being a node in an international system. It is technically quite possible to establish such systems.

Distributed implementation: A distributed system would be a system where pairs of institutions exchanged links, so that one database stores references to other databases or documents where persons in the internal system are also referenced. The distributed system may be more difficult technically to implement than the hierarchical, but it may be easier to implement organisational because it can be done in a bottom-up fashion.

We can do both! The hybrid model

The aim of this WG is that test systems will be developed using both of the methods described on the last slide. This is important in order to test different ways to solve the problem of co-reference. It is also quite possible to integrate systems designed by the two principles. Seen from the hierarchical side, a distributed co-reference system can be harvested to create a data warehouse of co-reference links with references back to the sources. Seen from a distributed system, the root node of a hierarchical system can be included as a large, but otherwise normal node in the distributed system.

Who will do the job?

There are two different entities that may do work in this area. Computer software is developed to detect and disambiguate names and other referring strings in texts. They are quick and reliable, but has a high level of mistakes: not-detected items, false positives, or both.

Human beings are quite reliable, especially in systems where they are working as professionals, e.g.cataloguers, curators or researchers. But they are quite expensive.

Information about the event in which the co-reference is stated should be stored. This makes it possible to differentiate between statements made by computer programs and statements made by persons, as well as between different persons. Depending on the need of the user, different levels of quality may or may not be included in the data set being investigated, and the different types of co-references are shown in result sets.

But how do we get people to do the extra effort?

A very important fact, and a main reason why we think this is possible to implement, is that the culture heritage workers are doing co-reference detection as part of their work today. We do not need to hire people to do something they do not do today. We just need to create systems (computational and organisational) so that they will be able to store the facts they detect in a way which may be usable in a co-reference system. Some cultural heritage systems have this option available today, others may develop it or connect external systems for it.

So there should ideally be no or very little extra effort for the people entering the data.

Use cases

There are many interesting use cases. One of them is the use within the institution to which the people who enter the data is connected. This may be as part of their further work on reference and co-reference detection, where growing co-reference systems will be of great help.

Another example is historical research, where co-reference systems will be very helpful in network analysis, using techniques such as friend-of-a-friend.

The co-reference working group

The CIDOC Co-reference Working Group was established at the CIDOC annual conference in Vienna in 2007. In the initial year, work have concentrated on two issues:

1. Monitoring of co-reference work being done in different parts of the world
2. Initiating co-reference projects, especially across national borders