AcroDef: A Quality Measure for Discriminating Expansions of Ambiguous Acronyms

Mathieu Roche and Violaine Prince

LIRMM - UMR 5506, CNRS, Univ. Montpellier 2,
34392 Montpellier Cedex 5 - France

Abstract. This paper presents a set of quality measures to determine the choice of the best expansion for an acronym not defined in the Web page. The method uses statistics computed on Web pages to determine the appropriate expansion. Measures are context-based and rely on the assumption that the most frequent words in the page are related semantically or lexically to the acronym expansion.

1 Introduction

Named Entities Recognition (NER) has become one of the major issues in Information Retrieval (IR), knowledge extraction from texts, classification, question answering (QA), and machine aided translation (MT). The state-of-the art literature in NER mostly focuses on proper names, temporal information, specific expressions in some technical or scientific fields for domain ontologies building, and so forth. A lot of work has been done on the subject, among which on acronyms, seen as particular named entities. Acronyms are very widely used in every type of text, and therefore have to be considered as a research issue as linguistic objects and as named entities.

An **acronym** is composed from the first letters of a set of words, written in uppercase style. This set of words is generally frequently addressed, which explains the need for a shortcut. It is also a specific multiword expression, such as "named entities recognition", abbreviated into NER, sometimes completely domain dependent (as NER or NLP are) and sometimes becoming a commonly used item (such as SARS, AIDS, USA, etc.). In some cases, acronyms become proper names referring to countries or companies (like USA or IBM). However, most of the time, acronyms are domain or period dependent. They are contracted forms of multiword expressions where words might belong to the common language. As contracted forms, they might be highly ambiguous since they are created out of words first letters. For instance NER, the acronym we use for `Named Entities Recognition` might also represent `Nippon Electrical Resources` or `Natural Environment Restoration`. Those are two other possible expansions for the acronym NER. An **expansion** is the set of words that defines the acronym. The word **definition** will also be used as a synonym for expansion in this context.

In all cases, an acronym behaves like a named entity. However, the intrinsic ambiguity in most acronyms enhances the difficulty of finding which exact entity

B. Kokinov et al. (Eds.): CONTEXT 2007, LNAI 4635, pp. 411–424, 2007.

is referred by this artificial name. Literature has been addressing acronym building and expansion (see section "state-of-the art") when the acronym definition is given in the text. However, choosing the right expansion for a given acronym in a given document, if no previous definition has been provided in the text, is an issue definitely belonging to NER, and not yet exhaustively tackled. The difficulty in acronym disambiguation is to automatically choose, as an expansion, the most appropriate set of words. This article tries to deal with this issue by offering a **quality measure** for each candidate expansion. In this context, let us name a a given acronym. For every a which expansion is lacking in a document d, we consider a list of n possible expansions for a: $a^1 ... a^n$. For instance, if IR is the acronym at stake, we could have IR^1= `Information Retrieval`, and IR^2 =`Investor Relations` (in finance and communication), and IR^3 =`Infra Red` (in optics and medicine). In a multilingual context, things could become worse, IR^4 = `Impôt sur le Revenu` (the French expression for income tax). Some web resources exist for providing acronym definitions (as an example, we use the site http://www.sigles.net/, which browses more than 17,000 sites in 212 countries).

The aim of our approach is to determine k ($k \in [1, n]$) such that a^k is the relevant expansion of a in the document d. To make such a choice, we provide a quality measure, called $AcroDef$, which relies on Web resources. The figure 1 summarizes the applied global process. The presentation is structured as following: section 2 discusses the output of the related literature, section 3 focuses on the quality measure $AcroDef$, where context and web resources are essential characteristics to be taken into account. Section 4 describes some experiments and discusses their results and finally conclusion and perspectives are suggested in 5.

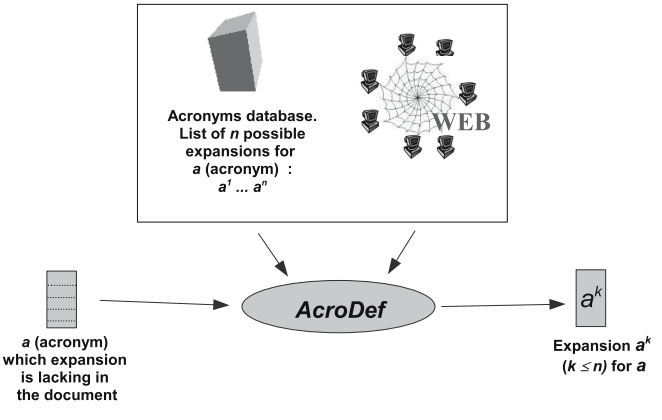

Fig. 1. Global process

2 Acronym Expansion Relevant Literature

Among the several existing methods for acronyms detection and expansion in literature, we present here some significant works. First, acronyms detection