Proper Name Classification in an Information Extraction Toolset

Peter Wallis, Edmund Yuen, and **Greg Chase**
Information Technology Division, DSTO, Box 1500 Salisbury, Sth Aust. 5108
{Peter.Wallis,Edmund.Yuen,Greg.Chase}@dsto.defence.gov.au

Abstract

Applied discourse analysis is a hot topic in Information Retrieval (IR) and the related field of Information Extraction (IE). Although interesting observations about discourse can be made "by hand," applications require large quantities of data about language — data which is rather uninteresting. This paper investigates using statistical analysis over a body of text to suggest new rules for recognizing named entities.

1 Introduction

Understanding human languages on any sort of scale is a knowledge intensive task. This paper describes a corpus based approach to gathering language data in the shallow parts of the NLP pond. Information retrieval is a popular application for researchers interested in applied NLP, but the problem of improving retrieval effectiveness appears to be intractable (Smeaton, 1992; Wallis, 1995). One helpful technique is tagging the proper names in text. Tagging and classifying (e.g. Is "Washington" a place or a person?) the named entities and co-references to them (she, he, the company) in text is also a primary concern in systems for *information extraction* (DARPA, 1995).

Information extraction (IE) is a well defined task; the aim being to extract data from free text, and put it in a more structured format. The IE task is not only well defined, it has application and is hence often seen as a prime example of language engineering, where the aim is to explicitly solve a problem rather than to understand the nature of language. IE systems have typically only been successful in narrow domains with significant effort required to move and existing information extraction system to a new problem domain. One approach is to use tools in a development environment that assists the language engineer to create a new information extraction system from pre-exisiting components.

The DSTO **Fact Extractor Workbench** provides the tools to create re-usable text skimming components, called fact extractors, that perform IE on a (very) limited domain. These components can be used directly to find things like dates and the names of companies including co-references, or they can be assembled to create larger fact extractors that skim text for more abstract entities such as company mergers.

The workbench provides different views of the domain text to assist in the development process. As an example, the language engineer might be interested in seeing how the word "bought" is used in the domain of interest. A "grep"-like tool allows him or her to view all and only those sentences containing "bought". Naturally more complex patterns are possible incorporating previously developed fact extractors in the pattern.

This paper discusses an extension to the corpus viewing tool set that assists the language engineer to find words, called *selector terms*, that may aid in the classification of proper nouns and determination of possible co-references for those nouns. First, we describe the domain in which we are applying our fact extractors. Next, we introduce our method of measuring the suitability of words as selector terms. Lastly we discuss how this data is collected and presented in the fact extractor workbench.

2 Problem Domain

The Named Entity Test is one component of the message understanding conference (MUC 5–7 (DARPA, 1995)) evaluations. The goal of the NE test is to add SGML tags to the evaluation texts that mark up all the proper names. The body of text used in these trials is a selection of articles from the Wall Street Journal. McDonald (McDonald, 1996) characterizes the problem as having three sub-components:

Peter Wallis, Edmund Yuen and Greg Chase (1998) **Proper Name Classification in an Information Extraction Toolset.** In D.M.W. Powers (ed.) *NeMLaP3/CoNLL98: New Methods in Language Processing and Computational Natural Language Learning,* ACL, pp 161-162.

- *delimit* the sequence of words that make up the name, i.e. identify its boundaries;
- *classify* the resulting constituent based on the kind of individual it names (e.g. Person, Organization, Location); and
- *record* the name and the individual it denotes in the discourse model

The emphasis in this paper is on a method for classifying the name using external evidence.

2.1 Classification

During this process, **internal evidence** (McDonald, 1996) may be gleaned as to the type of the named entity. Titles such as *Mr, Ms, Dr, Sir,* and *Jr* provide evidence of the named entity being a person. The presence of *Ltd.* or *G.m.b.H.* signify a company.

External evidence (McDonald, 1996) about a named entity's type can also be used. If it is unclear whether a name refers to a person or a company, it can help to look at the verb it participates in, or at any modifiers it may have. People do things like "head" organizations, "speak" and "walk". Companies "merge" and "take measures". People have employment roles, gender, and age; companies have locations and managing directors. Ideally a system would have rules that say if a subject-of-a-verb($< NE >$, (head, say, explain ...)) then the named entity is of type person. Similarly a function modified-by($< NE >$, (chairman, head, $< number >$ years old, ...)) could be used in a rule to determine if the $< NE >$ is a person. Writing such rules require a list of terms which are good **selector terms** for the entity of interest. The proposal is to add a tool to the fact extractor workbench that helps the language engineer find good selector terms using probabilistic measures.

3 Finding Class Selectors

To measure how good a selector term is for an existing fact extractor, we need to compare the problity that the word is present in a sentence and the probability that the word is in a sentence given that a "fact" is in that sentence.

w = word
S = sentence
S_f = sentence with fact f

$$Prob(w\ in\ S \mid f\ in\ S) = \frac{number\ of\ S_f\ with\ w}{number\ of\ S_f} \quad (1)$$

$$Prob(w\ in\ S) = \frac{number\ of\ S\ with\ w}{number\ of\ S} \quad (2)$$

If w and f are independent then 1 will approximate 2 however if they are dependent 1 will be different from 2.

A measure of w's selective power can be calculated as a ratio.

$$Sel_f(w) = \frac{Prob(w\ in\ S \mid f\ in\ S)}{Prob(w\ in\ S)} \quad (3)$$

An Sel of close to 1 indicates little correlation between the term, w, and the fact, f. An Sel significantly greater than 1 indicates a high degree of correlation between w and f and hence w is a good selector term. Interestingly, a Sel of significantly less than 1 (close to zero) indicates that the presence of w is a good indication of f being absent.

4 Incorporating Selective Power

A tool has been incorporated into the Fact Extractor Workbench that allows the user to run one or more fact extractors over the text corpus and produce and ordered set of candidate selector terms. This list of selector terms can then be considered for inclusion into a more refined fact extractor.

For example, by measuring the selective power of corpus words for the "City" fact extractor pattern, we can find which words are used in the context of Washington, the city and which are used in the context of Washington, the person. By ranking corpus words based on selective power, we single out candidates as good selector terms to refine the "City" fact extractor.

References

Defense Advanced Research Projects Agency (DARPA). *Proceedings Sixth Message Understanding Conference (MUC-6)*, November 1995.

David D. McDonald. Internal and external evidence in the identification and semantic categorization of proper names. In Branimir Boguraev and James Pustejovsky, editors, *Corpus Processing for Lexical Acquisition*, pages 21–39. MIT Press, Cambridge, Mass, 1996. A Bradford Book.

Alan F. Smeaton. Progress in the application of natural language processing to information retrieval tasks. *The Computer Journal*, 35(3):268–278, 1992.

Peter Wallis. *Semantic Signatures for Information Retrieval.* PhD thesis, Faculty of Applied Science, R.M.I.T., 1995.