自然言語処理 115－9
情報学基礎 43－1
（1996. 9. 12）

英語新聞記事からの固有名詞自動抽出技術

若尾 孝博

シェフィールド大学コンピューター・サイエンス学部

[滞在先] 日本電気（株）情報メディア研究所　音声言語研究部
〒216神奈川県川崎市宮前区宮崎4丁目1－1
Tel : 044-856-2152　Email : *wakao@hum.cl.nec.co.jp*

１９９５年１１月のＭＵＣ６大会では、英語新聞記事からの固有名詞(人名、地名、組織名)を自動的に高い精度で抽出出来ることが示された。その抽出技術の紹介として、シェフィールド大学の情報抽出システムにおける固有名詞の自動抽出の方法を説明する。このシステムの特徴は固有名詞を認定するあたり、多種類の情報を活用している点にある。統語的な情報だけでなく、意味論、語用論のレベルまでの情報、また固有名詞間での照応関係の情報などが活用されている。システムの評価として、各種の情報が固有名詞の認定にどれだけ貢献しているかを分析する。

Proper Name Extraction from English Newspaper Articles

Takahiro Wakao

University of Sheffield, Computer Science Department

[visiting] NEC Corporation, Information Technology Research Laboratories,
Human Language Research Laboratory
1-1, Miyazaki 4-Chome Miyamae-ku, Kawasaki, Kanagawa, 216, Japan
Tel : 044-856-2152　Email : *wakao@hum.cl.nec.co.jp*

One of the tasks at the latest Message Understanding Conference (MUC-6, November 1995) was to extract proper names from newswire text. Several systems showed good scores at their evaluation of the MUC convention. We will illustrate how, one of such systems, University of Sheffield information extraction system, works for extracting proper names from English newspaper articles.

1 はじめに

近年新聞記事など大量のテキストが電子化されてきているのに伴い、電子化テキストから重要な情報だけを抽出したいという要求が高まって来ている。英語新聞記事から、テキスト中の事実および事実間の関係を自動的に取り出して来る作業(以下では情報抽出と言う)は、米国国防省のＡＲＰＡ (Advanced Research Projects Agency) により運営されて来ているＭＵＣ (Message Understanding Conference) にて盛んに研究されている ([1] [2] [3])。

最近の大会である、the sixth Message Understanding Conference (MUC-6)は、１９９５年１１月に開催された。作業内容が今までのテンプレートを埋めるだけと言う作業から、４つの作業に細分化された。その４つの作業は以下の通りである：

- Named Entity
  固有名詞(組織名、人名、地名)、
  時間表現(曜日、年月日等)、
  数表現(金額、パーセント)の認定
- Coreference
  照応関係の認定
- Template Element
  ＭＵＣ５のテンプレートより規模を縮小
  組織名、人名、製品名の抽出
- Scenario Template
  シナリオが渡され、それによって指定される情報だけを抽出し、テンプレートを埋める

第１の作業は新聞記事中の固有名詞を自動的に認定し区分する Named Entity Task であり、その他の作業の基礎をなす基本的な作業だと位置付けられる。この作業には１５のグループが参加し、その内の約半数が評価の尺度であるRecall （再現率）とPrecision （適合率）の両方において９０％以上を記録した。人手で同じ作業をすると再現率、適合率とも平均で９５～９６％であることを考えると、かなりの好成績であった。対象は新聞記事の中の人事異動に関する記事で、限られた分野ではあったが、英語新聞記事からの固有名詞を自動的に高い精度で抽出する技術が開発出来たことが示された。

以下の章では、まずこの Named Entity Task を説明し、その後、参加グループの１つである、英国シェフィールド大学の情報抽出システム *LaSIE* を概説し、実際にシステムで使われた英語固有名詞抽出のための技術を具体的に解説する。またシステムの評価をモジュールごとに詳しく行った結果を示す。

2 Named Entity Task

Named Entity Taskでは新聞記事中の Named Entity (固有名詞、時間表現、数表現) を認定し、テキストに直接タグを付けるものである。タグはＳＧＭＬ(Standard Generalized Markup Language)タグである。Named Entity の分類は表１の通りである。

Class	Type	Exmaples
固有名詞	組織名	General Electric, GE
	人名	John Major, Clinton
	地名	Tokyo, U.S.A
時間表現	時刻	12 a.m.
	日付	July 18, Friday
数表現	金額	1 million dollars
	パーセント	90.5 %

表 1: ＭＵＣ６での Named Entity の分類

組織名 (organization) は会社名、政府組織機関、国際的機関名などを含む。人名 (person) は記事中の人名を認定するものであるが、人名に付いたタイトル (President, Mr., Dr. など) は含まない。地名 (location) は基本的に大陸、地域、国、州、市の名前である。

時間表現 (time expression) としては、日付、曜日、季節、会計年度などであり、数表現 (number expression) としては通貨単位を含む金額、パーセント(%)のついた表現が抽出の対象となった。

3 LaSIE system

LaSIE システムは英国シェフィールド大学にて開発された情報抽出システムであり、ＭＵＣ６の Named Entity Task を含む４つの作業に参加した。システムの概略は図１の通りである。

本システムは基本的にパイプライン構造で、３つのモジュール、lexical preprocessing, parsing, discourse interpretation から構成されている。各モジュールの要約は次の通りである。

- lexical preprocessing
  入力されたテキストを読み込み、入力文を「トークン」、つまり、個々の単語及び句読点に分割する。次に、形態素解析が行われ、品詞情報がトークンに付けられる。固有名詞および固有名詞の存在を示すキーワードがマークされる。その後、チャートパーサー

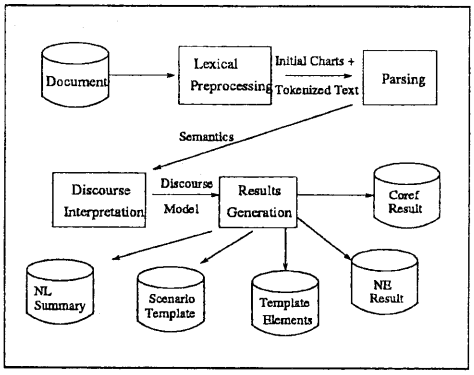

図 1: LaSIE システムの概略

のための初期チャートが生成され、次のパーサー(構文解析)モジュールに渡される。

- parsing
  構文解析は 2 回行われる。初回は固有名詞専用の文法ルールを使い、2 回目は文章レベルの文法ルールが使われる。構文解析を行うのと同時にその文章の意味表現 (semantic representation) も作成される。
- discourse interpretation
  作成された意味表現をシステムの持つ世界知識(world model, 一般知識を階層的に表したもの)に反映させて、照応関係の認定やディスコースレベルでの推論を行う。

固有名詞の抽出に関して、*LaSIE* 情報抽出システムの特徴は、種々のレベルでの情報、形態素、統語的、意味的、および語用論のレベルでの情報を総合的に使って固有名詞を認定し、区分しているところである。つまり、固有名詞の抽出はシステム全体で行なわれ、専用のモジュールは存在しない。このアプローチは「総合的アプローチ」 (comprehensive approach) と呼ばれ、次の 4 つがその柱となっている。

1. リストの参照
   各種の固有名詞の名前のリスト、および固有名詞の存在を示すキーとなる語 (trigger words) のリストを使用。
2. 固有名詞専用の文法
   構文解析をする際にまず固有名詞専用の文法を用い、その後に文章レベルの文法が使われて、二度目の構文解析が行われる。二度目の構文解析の時点では、初回の解析で固有名詞と認定されたものは、トークンが複数であっても 1 個のトークンとまとめて取り扱われる。
3. 固有名詞ための照応関係の認定
   組織名、人名はテキストの初めに正式名称で現れ、その後は短縮された形となることがよくある。これに対処するために固有名詞ための照応関係を、一般の照応関係認定とは別に特別に行う。
4. ディスコースレベルでの推論
   構文解析と同時に作られた意味表現をシステムの持つ世界モデルに反映させて、固有名詞獲得のための推論を行う。

尚、システムの詳細は [5] を参照して頂きたい。

4 固有名詞抽出の技術

ここでは、固有名詞抽出ための具体的な方法・技術を *LaSIE* システムのモジュールを追って説明して行く。

4.1 Lexical preprocessing

インプットされたテキストは、まず個別のトークンに分割され、Brill's tagger ([6]) にて品詞が付けられる。品詞は基本的に Penn Treebank で使われている 4 8 個と特別に今回使用した(例えば時間表現用にdate, time) 数個である。大文字で始まる語は、文頭を除き、固有名詞(またはその一部)と仮定する。つまり、Brill's tagger のデフォルト設定で、タガーの辞書になく、大文字で始まる語は固有名詞を示す、`NNP` (単数)、`NNPS` (複数)の品詞タグが付く。次に、固有名詞を集めたリストととの参照が行われ、テキスト中のトークンでリストにある語・語句は固有名詞であるという情報が付け加えられる。使用されたリストは次の通りである。

- 組織名：約 2 6 0 0 の企業名と政府関係機関名、MUC-5 で使用された Wall Street Journal の記事より半自動的に作られたもの。
- 地名：約 2 2 0 0 の国、州(province, state, など)、市名。
- 人名：約 5 0 0、 Oxford Advanced Learner's Dictionary ([7]) に載る人名 (first name)。

これらのリストに加えて、固有名詞の存在を示すキーとなる語 (trigger words) も参照が行われ、

トークンと同じものがあればその情報が付加される。trigger word のリストは次の通りである。

- 企業名接尾語 (company designators) : 'Co.', 'PLC' など９４個。
- 企業名を示す語 : 'Association', 'Bank' など１３５個。
- 航空会社を示す語 : ３個、'Airlines', 'Airways', 'Air'
- 政府関係機関を示す語 : 'Court of Appeals', 'Legislature' など7個。
- 人名接頭語 (human titles) : 'President', 'Mr.' など約１６０個。
- 地名を示す語 : 'Gulf', 'Mountain' など８個。

これらの trigger words は基本的に手作業で集められたが、「企業名を示す語」だけは半自動的に収集された。MUC-6 の開発用の新聞記事は約２００あり、企業名が既にマークされている。その記事中の企業名を取り出し、'of' を含まない企業名の最後に来る語、及び、'of' を含む企業名の'of' 直前の語を集めて、整理をしてリストとした。。例えば 'Rockwell International' の 'International' や 'Association of Air Flight Attendants' の 'Association' などである。

4.2 Parsing

LaSIE システムのパーザーは、Prolog で書かれたボトムアップのチャートパーザーである。使われた文法はフィーチャーを基本とした(feature-based) 文脈自由文法である。構文解析を行うと同時に意味表現(semantic representation) がユニフィケーションを使って構築される。構文解析の終了時に、解析された構文が複数ある時は最適構文が一つ選択される。構築された意味表現は、次のディスコース分析の基礎となる。

構文解析は２つの別々の文法規則を用いて２段階に分けて行われる。第一段階で固有名詞専用の文法が用いられてテキスト中の固有名詞が認定され、第二段階で文レベル文法を使っての解析が行われる。第二段階では、第一段階で認定された固有名詞は、複数トークンの名前であっても、１つのトークンとして扱われる。

4.2.1 固有名詞専用文法

固有名詞のための文法は全て名詞句を作るもので、文レベル文法の名詞句規則の一部と見なすことが出来る。規則数は１７７個で、９４個が企業名用、５４個が人名、１１個が地名用である。企業名用規則の一部を次に示す。

```
NP --> ORGAN_NP
ORGAN_NP --> LIST_LOC_NP NAMES_NP CDG_NP
ORGAN_NP --> LIST_ORGAN_NP NAMES_NP CDG_NP
ORGAN_NP --> NAMES_NP '&' NAMES_NP
```

非終端記号 LIST_LOC_NP, LIST_ORGAN_NP と CDG_NP はリスト参照の段階で付けられたもので、リスト中にあった地名、組織名、企業名接尾語を示している。また、NAMES_NP は未分類固有名詞を示している。ORGAN_NP --> NAMES_NP '&' NAMES_NP は、未分類固有名詞の次が '&' で、その次にまた未分類固有名詞が来る場合は、それを組織名とすると言う規則である。例えば、'Marks & Spencer' や 'American Telephone & Telegraph' などがこの規則により、組織名となる。

```
LOCATION_NP -->
  NAMES_NP COMMA TAGGED_LOCATION_NP(PROVINCE)
```

地名の規則の一部であるが、同格 (appotition) の関係にある地名を認定するために規則である。TAGGED_LOCATION_NP はリスト参照の段階で地名とされたものである。第一の規則によると、未分類固有名詞の次にコンマが来て、その直後にリスト参照により既に分かっている地名が続く場合は、これ全体が地名となる。この時構築される意味表現には、最初の未分類固有名詞が地名であると記されるようにしている。例えば 'Fort Lauderdale, Fla.' において、'Fla.'が地名であると分かっている場合、'Fort Lauderdale' も地名とすることになる。

最も構造が複雑な組織名に関する規則が一番多く、人名については人名接頭語や人名中の特別な語 ('J. Ignacio Lopez de Arriortua' 中の 'de'など)を扱う規則があるため、規則数が二番目に多くなっている。

4.2.2 文レベルの文法と意味表現

文レベルの文法は規則数が約１１０で、Penn TreeBank-II ([8], [9]) から自動的に抽出されたものである([10])。構文解析と同時に構築される意味表現を生成する規則は、手作業で作成され、各文法規則に割り当てられた。固有名詞には関しては、それまでに分かった情報に加えて、name 属性が設けられて、固有名詞の文字列をその値として代入している。例えば 'Ford Motor Co.' が組織名として既に認定されている場合、その意味表

現は、company(e23) & name(e23,'Ford Motor Co.') と言ったものになる。

4.3 Discourse interpretation

ディスコースレベルにおいては、構文解析と同時に作られた意味表現をシステムの持つ世界モデルに反映させて、ディスコース単位での意味表現を構築している。この世界モデルは、世界に関する一般的な知識ではなく、むしろ人事異動と言う限られた分野での知識、つまり人や組織に関する知識を持つものである。そこで、固有名詞獲得のために2つの事が行われる。第一が固有名詞のための照応関係の認定であり、第二が意味表現において特定の関係にあるエンテティー (entity) についての推論を行うことである。

4.3.1 固有名詞用の照応関係認定

固有名詞のための照応関係認定 (coreference resolution) は、固有名詞、特に、組織名のバリエーションを確実に認定するためである。例えば、テキスト中に 'Ford Motor Co.' がまず現れ、その後は 'Ford' が使われていて、しかも未分類固有名詞となっている場合に'Ford Motor Co.' と 'Ford' を同一の会社であると認定することにより、'Ford' が 'Ford Motor Co.' が持つ属性と同じ属性を持つことし、'Ford' が組織名であると認定する[1]。2つの固有名詞が与えられた場合に、その2つの名前が同一のものを示しているのかを、つまり照応関係にあるのかを調べるために、幾つかのヒーリスティックが用いられた。次にそのヒーリスティックの例を示す。

与えられた2つの名前を Name1 と Name2 として、

- もし、Name2 が Name1 の単語列の一部で、その順序も同じであるときは Name1 は Name2 とマッチする。
  例、'American Airlines Co.' と 'American Airlines'
- Name1 が人名で、Name2 がその first name か family name であるとき、Name1 は Name2 とマッチする。
  例、'John J. Major' と 'Major'

このようなヒーリスティックは、組織名用に31、人名用に11、地名用に3ある。組織名用のルールには省略形 ('IBM' と 'International Business Machine' など) を認定するものも含まれている。詳しくは [4]を参照して頂きたい。

4.3.2 意味表現を使った推論

固有名詞獲得のために、次のような場合に、意味表現での推論がなされる。

- 名詞と名詞間の修飾
  もし未分類の固有名詞が組織名に関連が深いものに修飾している場合、その未分類の固有名詞は組織名と認定される。
  例、'Erickson stocks' において、'stock' は組織名に関連が深いものと判断されて、'Erickson' が組織名と認定される。
- 所有格 (possessive)
  もし未分類の固有名詞が企業における役職名と所有の関係にある場合、その未分類の固有名詞は組織名と認定される。
  例、'vice president of ABC', 'ABC's vice president' など。

意味表現を使った推論は、固有名詞の周りの文脈から情報を獲得しようとするもので、McDonald ([11]) の言うところの external evidence (固有名詞文字列外からの情報) に相当する。

5 結果と評価

これらの処理の後、MUC-6 で指定された形、オリジナルのテキストに予め定められた SGML タグを挿入する形で最終結果が生成される。このシステムが生成した結果が人手で作られた「正解」と照らし合わされて、自動的に評価、再現率、適合率の算出、が行われる。

評価に使われたのは、システムにとって全く新しい Wall Street Journal の新聞記事30記事であった。30記事中に、449の組織名、373の人名、110の地名、111の時間表現があった。[2] *LaSIE* システムの結果に対する評価(再現率、適合率)は次の通りである[3]

[1] Named Entity Task では照応関係を示す必要がない。しかし、Coreference の作業では固有名詞間の照応関係を認定することは作業の一部である。

[2] 数表現もあったが、本論文の評価からは外している。

[3] 注：これらの数字は MUC-6 の大会で発表された数字とは少し異なっている。大会に提出された結果は、システムが実際に29の記事しか処理出来なかったものであり、ここに載せるものは、30記事全てを処理した結果である。

区分	再現率	適合率
組織名	91 %	91 %
人名	90 %	95 %
地名	88 %	89 %
時間表現	94 %	97 %
全体	91 %	93 %

表 2: *LaSIE* システムの成績

5.1 各モジュールの貢献度

システム全体を 4 段階に分けて、段階毎に結果を出し、各モジュールが固有名詞の抽出に対してどれだけの貢献をしているのかを調べてみた。

4 段階のセッティングは次の通りである。

- **setting 1** : lexical preprocessing だけが使われた場合で、品詞付けとリストの参照のみの結果である。
- **setting 2** : 構文解析が setting 1 に追加された場合。
- **setting 3** : setting 2 に固有名詞用の照応関係認定が加えられた場合。
- **setting 4** : フルのシステム、Discourse interpretation がフルになされた場合。

表 3 に示したのは、全固有名詞に対する各セッティングにおける成績である。今までの再現率 (Recall) と適合率 (Precision) に加えて、Van Rijsbergen ([12]) の F-measure (P&R) が載せてある。再現率と適合率の重みを同一として、F-measure は次の公式で算出される。

$$F = \frac{2 \times \text{適合率} \times \text{再現率}}{\text{適合率} + \text{再現率}}$$

セッティング	再現率	適合率	P&R
setting 1	49	89	63.01
setting 2	79	94	85.82
setting 3	89	94	91.13
setting 4	91	93	91.75

表 3: 各モジュールの貢献度

表 3 より、リスト参照と構文解析だけで、かなりの精度で固有名詞を抽出出来ることが分かるが、更に良い結果を出すには、固有名詞ための照応関係認定(setting 3) と更なる文脈からの情報 (setting 4) が必要であることが分かる。

5.2 固有名詞各クラスごとに見た各モジュールの貢献度

次に、固有名詞各クラス別(組織名、人名、地名、時間表現)についてモジュールの貢献度を見てみる。組織名に対する各モジュールの貢献度は表 4 ようになる。

セッティング	再現率	適合率	P&R
setting 1	46	87	59.91
setting 2	65	92	76.15
setting 3	87	93	89.84
setting 4	91	91	91.13

表 4: 組織名に対する各モジュールの貢献度

人名、地名、時間表現に対する各モジュールの貢献度は表 5 、 6 、 7 に示される。

セッティング	再現率	適合率	P&R
setting 1	47	88	61.64
setting 2	89	95	92.34
setting 3	90	95	92.14
setting 4	90	95	92.14

表 5: 人名に対する各モジュールの貢献度

セッティング	再現率	適合率	P&R
setting 1	81	94	86.84
setting 2	88	90	88.99
setting 3	88	89	88.58
setting 4	88	89	88.58

表 6: 地名に対する各モジュールの貢献度

セッティング	再現率	適合率	P&R
setting 1	32	100	48.97
setting 2	94	97	95.41
setting 3	94	97	95.41
setting 4	94	97	95.41

表 7: 時間表現に対する各モジュールの貢献度

これらの結果をグラフにまとめると、図 2 のようになる。これらの分析が示すように、setting 3 、setting 4 が効果を発揮するのは組織名の場合である。

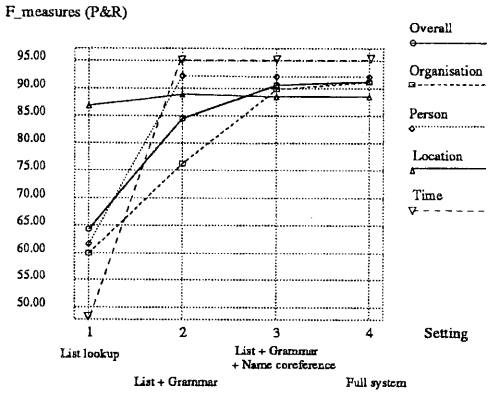

図 2: 固有名詞各クラス別に見た各モジュールの貢献度

6 結論

MUC-6 に参加したシェフィールド大学の情報抽出システム *LaSIE* で固有名詞の抽出に関してどのような技術が用いられたのかを説明した。そして、その抽出の結果に対する詳しい評価を行った。その結果として言えることは、次の通りである。

1. 固有名詞文字列内の情報に頼った抽出方法、つまり、リストの参照や構文解析だけで、ある程度までの成果がでるが、より良い成果を収めるには名前の周りにある文脈からの情報、特に、固有名詞の照応関係や意味のレベルでの推論が必要であることが分かった。特に組織名においては、この事がはっきりと現れた。

2. 人名、地名、時間表現については、固有名詞文字列内の情報に頼った抽出方法でも、高い適合率を保ちながら高い再現率を達成することが可能であることが分かった。つまり各モジュールの貢献度が固有名詞のクラスによって異なることが明確になった。

3. 今回対象となった新聞記事中の固有名詞の内４０％余りが組織名であることを考えると、固有名詞抽出において、高再現率と高適合率を達成するには、文脈からの情報をいかに獲得し、利用するかが重要であると分かった。

また、もう少し一般的に言って、２つの結論が引き出せると思われる。第一は、固有名詞抽出のシステムを評価する場合、単に全固有名詞に対する成果を分析するだけではなく、システムのモジュール毎、固有名詞のクラス毎に評価することにより、より詳しい評価が出来、システムの採用した方法のより詳しい検討が出来ることである。第二は、*LaSIE* で採用されたアプローチ、つまり、種々のレベルでの情報を総合的に使って固有名詞を認定し、区分する「総合的アプローチ」が適切なものであると言うことである。このアプローチを採用することにより、ただ単に新聞記事中の固有名詞の認定だけに留まらず、他の作業 (例えば、照応関係の認定、テンプレート埋め、要約の生成など) を助けることにつながると考える。

7 謝辞

本研究は英国 the Department of Trade and Industry (Grant Ref. YAE/8/5/1002) 及び、the Engineering and Physical Science Research Council (Grant # GR/K25267) からの研究費により可能となった。

参考文献

[1] *Proceedings of the sixth Message Understanding Conference (MUC-6)*, Columbia, Maryland, U.S.A. 1995, Morgan Kaufman Publishers Inc. 1996.

[2] Ralph Grishman and Beth Sundheim "Message Understanding Conference - 6: A Brief History" In *Proceedings of the 16th International Conference on Computational Linguistics (Coling 96)*, 1996.

[3] 若尾 孝博, "英語テキストからの情報抽出：MUC第６回大会の参加報告", 電子情報通信学会技術研究報告 NLC-96-9-20, 1996.

[4] Takahiro Wakao, "A Comprehensive Approach for Proper Name Recognition and Classification for Information Extraction", PhD Dissertation, University of Sheffield, Computer Science Department (to appear).

[5] R. Gaizauskas, T. Wakao, K. Humphreys, H. Cunningham, Y. Wilks "University of Sheffield: Description of LaSIE system as used for MUC-6" In *Proceedings of the Sixth Message Understanding Conference (MUC-6)*, November 1995, Columbia, Maryland, U.S.A., Morgan Kaufman Publishers Inc. 1996.

[6] E. Brill "Some Advances in Transformation Based Part of Speech Tagging" In *Proceedings of the Twelfth National Conference on Artificial Intelligence (AAAI-94)*, Seattle, Washington., 1994.

[7] A.S. Hornby (Ed.) *Oxford Advanced Learner's Dictionary of Current English*, Oxford University Press, London, 1980.

[8] M.P. Marcus, B. Santorini and M.A. Marcinkiewicz "Building a Large Annotated Corpus of English: The Penn Treebank", *Computational Linguistics*, 19 (2): 313–330, 1993.

[9] M. Marcus, G. Kim, M.A. Marcinkiewicz, R. MacIntyre, A. Bies, M. Ferguson, K.Katz and B. Schasberger "The Penn Treebank: Annotating Predicate Argument Structure Distributed on The Penn Treebank Release 2 CD-ROM ", by the Linguistic Data Consortium, 1995.

[10] Robert Gaizauskas, "Investigations into the Grammar Underlying the Penn Treebank II", Department of Computer Science, Univeristy of Sheffield, Research Memorandum, CS-95-25, 1995.

[11] D. D. McDonald "Internal and External Evidence in the Identification and Semantic Categorisation of Proper Names.", In *Proceedings of SIGLEX workshop on "Acquisition of Lexical Knowledge from Text"*, pp. 32–43, Ohio, U.S.A. 1993.

[12] C.J. Van Rijsbergen *Information Retrieval*, London: Butterworths, 1979.