An Efficient Information Extraction Model for personal named entity

Teena A.Sunny[1], G. Naveen Sundar[2]

[1,2]*Dept. of Computer Science, Karunya University*
India

Abstract: ***Named entity recognition (NER) is one of the key techniques in language processing tasks such as information extraction. This paper focuses mainly on recognition of named entity using distance based clustering and attributes extraction patterns. The ultimate goal of the paper is to reduce ambiguity of person names with higher precision and recall and to avoid duplicity.***

Keywords: ***Unsupervised learning, precision, recall, Ambiguity, Bigrams, attribute extraction, clustering, tokens***

I. INTRODUCTION

The rapid growth of the world-wide web has resulted in a huge amount of information sources on the Internet. However, due to the heterogeneity and the lack of structure of Web information sources, access to this huge collection of information has been limited to browsing and searching. In general information retrieval (IR), mainly concerns how to identify relevant web page from a collection of input web pages, has become crucial to many applications of Web mining and searching tools. In research the notion of personal information extraction has become vital for searching people on the web.

Information extraction (IE) is the task of automatically extracting structured information such as entities, relationships between entities, and attributes describing entities from unstructured or semi-structured machine readable documents.

This paper it is mainly a combinational model. It's an end to end system which has wider applications in the field of social networking. It provides solution to two basic issues i.e. ambiguity and duplicity. Practically it has two sub divisions namely disambiguation system and attribute extraction. The proposed system analysis shows that after web pages sharing common name are identified they are given as input for attributes extraction through various extraction patterns by page-pair wise comparison whereby most relevant web pages are identified through attribute mapping. This approach clearly states that the system achieves higher accuracy and reduces ambiguity by eliminating duplicate attributes of same person.

II. PREVIOUS WORK

The study of varying system states that there are mainly two kinds of features: document level features which refer to information retrieval from a given webpage whereas global features refer to information derived beyond the given corpus, such as information from an extra corpus and so on. Due to ambiguity of personal names, most previous disambiguation systems use unsupervised clustering as the basic approach, despite the different features used to create the similarity space. Besides the given corpus, there is a large amount of online information about the focus person. Since most of the given corpora include only a small part of documents containing the ambiguous personal names, it is sometimes hard to make a co-reference decision, even for an annotator. Because of the varying differences in the weights Web IT 5-gram was chosen as web corpus. In [9],[2] and [6] earlier systems have adopted various different clustering techniques like k-clustering, hybrid clustering, agglomerative clustering etc. to chose the most efficient and appropriate web page.

It is very well known that web corpus is heterogeneous and noisy. The common problem for AE is that they are designed only for news corpus, so handling noisy web data becomes tedious. Therefore, most of the current NER and RDR systems cannot work well because they heavily rely on surface cues, which are usually noisy in Web data. After experimental analysis as in [10], the earlier systems as in [1] and [8] have developed algorithms that segments a webpage into fragments according to their writing style: formal style and informal style, and then used a rule-based system, MT based AE and hybrid AE to catch this Web expression property, develop some patterns working on fragments which achieve high precision but lower recall. To catch this Web expression property, we develop some patterns working on fragments, and the experiment shows that those patterns could achieve high precision, which is very important for real applications.

This paper has adopted a different AE system which extracts the most relevant of all the attributes and then maps them to find the exact sets.

III. METHODOLOGY

A. *Dealing with feature extraction*

This methodology as in [1] improves the robustness by retrieving features using unsupervised approaches. For a given ambiguous personal name and a set of web pages including a mention of the query personal name, firstly need to define the object to be disambiguated. Each web page undergoes a simple pre-processing where the parser extracts a clean document. Then, features like token-based features, n-gram features, a snippet-based feature are extracted, and for each feature the cosine value is computed. Any of the four kinds of tokens are extracted which can appear in any part of the webpage i.e. either body or title etc.

- Query name tokens: the tokens occurring in sentences that include a mention of the focus personal name.
- Full tokens: the tokens occurring in a given web page.
- URL tokens: the tokens occurring in the corresponding URL of a given web page.
- Title tokens in root page: the tokens occurring in the title of the root page of a given web page.

For a substring (token) w in S, its TFIDF weight is computed as in Eqs. (1) and (2)

$$V(w,S) = \frac{V'(w,S)}{\sqrt{\sum_{w \in S} V'(w,S)^2}} \quad (1)$$

$$V'(w,S) = \log(TF_{w,s}+1) \times \log(IDF_w) \quad (2)$$

where $TF_{w,S}$ is the frequency of substring w in S and IDF_w is the inverse of the fraction of web pages in the given corpus that contains w.

But only extracted tokens cannot fully solve the issue so the weights of tokens are re-learned in a web page, and extract some meaningful bigrams. To learn the real weight for a token w, first the frequency of this token is obtained by searching token in the Web 1T 5- gram corpus, and then its weight is calculated by the above TFIDF formula. The snippet-based feature extracts the token information from snippets directly available from the search retrieval, which often includes more information about the ambiguous object beyond the given web page.

B. *Dealing with distance based clustering*

A document is represented with a vector just like a normal data point. Any clustering method has to embed the documents in a suitable similarity space. It is common to use similarity to compare two documents. The distance between one cluster and another cluster to be equal to the shortest distance from any member of one cluster to any member of the other cluster. If the document consists of similarities, system considers the similarity between one cluster and another cluster to be equal to the greatest similarity as in [4]. Fig. 1 shows how feature extraction and clustering combine to become a disambiguation system.

In this paper the most commonly known similarity function namely the cosine similarity metric is used. Cosine similarity is a measure of similarity between two vectors of an inner product space that measures the cosine of the angle between them. For given two vectors A and B, the cosine similarity'θ' is represented using a dot product and magnitude. In the case of information retrieval, the cosine similarity of two documents will range from 0 to 1. Thus based on the similarity score the web pages related to the same query name is gathered which filters out the other web pages. This allows documents with the same composition, but different total to be treated identically which makes this the most popular measure for text documents.

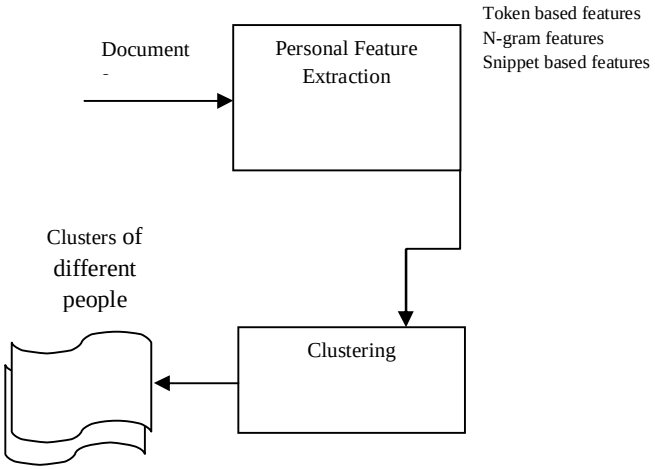

Fig. 1 Flow Diagram of Person Disambiguation

C. *Dealing with attribute extraction*

Adopting processing pipelines with multiple techniques including named entities recognition, regular expression patterns and gazetteer based matching; manually constructed rules based on cleaning. Attribute extraction system must be divided into two steps. First, AE should identify the attribute from web pages as listed in Table 1 using gazetteer or matching regular expression as in [3] and then retrieve attribute values for a prelisted attributes thus decide which among is relevant by finding its proximity.

For example, if the given name is John Fitzgerald Kennedy then this process will generate variants such as J. F. Kennedy, John F. Kennedy, Kennedy J. F., and John Kennedy. The system generate the variants like: first name followed by last name, last name followed by first name, a comma appearing between the two names, a word appearing

between the two names, first name initialized and immediately followed by the last name, last name followed by a comma and the first name initialized, and first name initialized and followed by a word and the last name. To find the attributes of the given person, we find the distance for each marked attribute value from a name variant. It then select the closest attribute value as the correct candidate.

To mark the potential values of attributes we use lists of candidate attribute values and a set of manually created rules in the form of regular expressions.

- Date of birth: A set of rules in the form of regular expression is used to mark all date strings in the text. Then normalize all date strings to YEAR/MONTH/DAY format.
- Email: E-mail addresses are also marked using the regular expression.
- Occupation: A list of occupations is created from different sources. Moreover tokenized each entry in the occupation list into words and sorted the words according to their total frequency within the list. The goal of this is to identify words that are commonly used to describe occupations. If a sequence of words contains any of those high frequency words, system selects those sequences as occupations.
- Award: Likewise used different sources to create a list of awards. Any entry that is found in this list is marked as an occurrence of an award in the given text. The system performs a word frequency analysis on this list and finds the most commonly used words in names of awards.

TABLE 1

ATTRIBUTE VALUE AND THEIR TYPES FOR EXTRACTION METHOD

Category	**Attributes**
Typical pattern	Phone Fax Email Date of birth
NER with limited candidates	Degree Nationality Occupation
Traditional NER	Birth place Relatives Other name

The advantage of using regular expression is that it's more flexible, portable and scalable, and potentially having higher precision and recall. It also has the advantage of being applicable to new languages for which no developer with sufficient knowledge of the language is available. Another way adopted is to construct initial database (gazetteer) of different occupation and awards patterns.

Then for each webpage for the matching attribute's candidate values the proximity is found which concludes the exact attribute of the person in query and assigns the exact attribute. Thus attribute extraction is done on cluster of the selected web pages only.

D. Dealing with page-pair wise mapping

Finally for each individual web page the proposed system extracts the attributes and their candidate values. But the major issue still arising is that the same candidate values can exist in inconsistent text formats across web pages, making it difficult to identify matching objects using exact text match. Therefore in order to find the matching objects the system performs comparison of the shared attribute values as in [5]. Moreover the system performs possible mappings between objects of each web page.

When comparing objects, the alignment of the attributes is determined by the user. The values for each attribute are compared individually (e.g. Name with Name, DOB with DOB, and Email with Email).

The mapping learner determines which most relevant attribute, or combinations of attributes for mapping web pages. The purpose is to achieve the highest possible accuracy various application domains. In this approach, the system actively chooses the most informative candidate mappings for the user to classify as mapped or not mapped in order to minimize.

Thus mapping rules determines which information about attributes is important for determining the mapping between pair of web pages.

IV. CONCLUSION

This paper clearly states that system has tried to identify both common and uncommon attributes of similar pages referring to the same person. Profoundly system reduces the ambiguity by eliminating duplicate attribute values for the same person and scores a higher accuracy with high precision and recall. Thereby suppresses the most raised issue of the earlier personal name disambiguation system.

REFERENCES

[1] Ying Chen, Sophia Yat Mei Lee, Chu-Ren Huang, "*A robust web personal name information extraction system*", In the proceedings of Expert Systems with Applications 39 2690–2699 (2012).

[2] Mann, G. & Yarowsky, D., "*Unsupervised Personal name disambiguation*", CoNLL (2003).

[3] Watanabe, K., Bollegala, D., Matsuo, Y. & Ishizuka M., "*A two-step approach to extracting attributes for people on the web in web search results*", In 18th www conference 2nd web people search evaluation workshop (WePS 2009).

[4] Alexander Strehl, Joydeep Ghosh, and Raymond Mooney, "*Impact of Similarity Measures on Web-page Clustering*", In the proceedings of AAAI Technical Report WS-00-01.

[5] Sheila Tejada, Craig A. Knoblock, Steven Minton, "*Learning Domain Independent String Transformation Weights for High Accuracy Object Identification*", In proceedings of SIGKDD 2002.

[6] Artiles, Javier, Julio Gonzalo, Satoshi Sekine,"*WePS 2 evaluation campaign: Overview of the web people search clustering task*", In 2nd web people search evaluation workshop, 18th www conference,(2009).

[7] Pedersen Ted, Kulkarn, nagha, "*Unsupervised discrimination of person names in web contexts*", In Proceedings of the eighth international conference on intelligent text processing and computational linguistics, Mexico City, 2007.

[8] Sekine, Satoshi & Artiles, Javier, "*WePS 2 evaluation campaign: overview of the web people search attribute extraction task*", In 18th www conference 2nd web people search evaluation workshop (WePS 2009).

[9] G. Naveen Sundar, Teena A. Sunny, "*Optimized Clustering for Personal Name Information Extraction*", In the proceedings of IJCST 2012.

[10] Javier Artiles, Julio Gonzalo, Satoshi Sekine "The SemEval-2007WePS Evaluation: Establishing a benchmark for the Web People Search Task", In the Proceedings of the 4th International Workshop on Semantic Evaluations, pages 64–69 (SemEval 2007)