UNN-WePS: Web Person Search using co-Present Names and Lexical Chains

Jeremy Ellman
Northumbria University
Pandon Building
Newcastle upon Tyne
UK
Jeremy. Ellman @unn.ac.uk

Gary Emery
Northumbria University
Pandon Building
Newcastle upon Tyne
UK
Gary.Emery@unn.ac.uk

Abstract

We describe a system, UNN-WePS for identifying individuals from web pages using data from Semeval Task 13. Our system is based on using co-presence of person names to form seed clusters. These are then extended with pages that are deemed conceptually similar based on a lexical chaining analysis computed using Roget's thesaurus. Finally, a single link hierarchical agglomerative clustering algorithm merges the enhanced clusters for individual entity recognition. UNN-WePS achieved an average purity of 0.6, and inverse purity of 0.73.

1 Introduction

Guha and Garg (2004) report that approximately 4% of internet searches are to locate named individuals. Yet, many people share the same name with for example 157630 individuals in the UK sharing the most common name 'David Jones' (UK statistics cited by Ash 2006). Consequently identifying web pages on specific individuals is a significant problem that will grow as everyone acquires a web presence.

There are several proposed approaches to identifying which individuals correspond to which web pages. For example, Bollegala et al. (2007) propose augmenting queries in the style of relevance feedback (Salton and Buckley 1990), Kalashnikov (2007) treat Web Person Search (WePS) as a disambiguation problem whose objective is to distinguish individuals, whilst Wan et al. (2005) see WePS as a clustering problem.

WePS has both similarities and differences to word sense disambiguation (WSD). Both seek to classify instances of usage, but in WSD the sense inventory is fixed. WSD then is more amenable to a classification solution where a system can be effectively trained using learning algorithms. In WePS we do not know from the outset how many individuals our pages correspond to. Consequently we took the view that WePS is better seen as a clustering rather than a classification problem.

1.1 Ambiguity

Ambiguity is a common feature of WePS and WSD. There are multiple types of ambiguity in the relation between person names and entities that confound overly simple approaches. Firstly, note that some first names are also last names (Les Paul, Michael Howard), and that some last names also occur as given names (Woodrow Wilson Guthrie, Martin Luther King). Consequently, an overly simple name parser will easily be confused. Secondly many last names are also place names (Jack London, William Manchester). Thus, if a last name is not found in the names database, but is found in the gazetteer, a name can be confused with a location. Finally, we come to toponym ambiguity, where the name of a place may correspond to several locations. (For example, there are thirteen places called Manchester, multiple Londons, Washingtons etc.) Resolving toponyms is a research problem itself (Leidner, 2004).

1.2 Statistics

Statistics are a further relation between WePS and WSD. We expect Zipf's law (e.g. Adamic and Huberman 2002) to apply to the relation between

Proceedings of the 4th International Workshop on Semantic Evaluations (SemEval-2007), pages 402–405,
Prague, June 2007. ©2007 Association for Computational Linguistics

web pages and individuals, meaning that relative frequency and rank form a harmonic series. In other words some people will be associated with many pages and increasingly more will be linked to fewer. This has a strong link to disambiguation, where an inaccurate algorithm may give inferior performance to the strategy of always selecting the most frequent sense.

Now if we consider the types of data that distinguish individuals, we might find colleagues, friends, and family mentioned in web pages, in addition to locations, dates, and topics of interest. Of these, names are particularly useful, and we define co-present names as names found in a web page in addition to the name for which we are searching.

Names are statistically useful, even though many people share the same name. For example there are 7640 individuals in the UK (for example) that share the most popular female name "Margaret Smith". Given the population of the UK is approximately 60 million, the probability of even the most common female name in the UK occurring randomly is 1.27+10-4 (of course not all the individuals have web pages).

Now, Semeval WePS pages (Artiles 2007) have been retrieved in response to a search for one name. Often such web pages will contain additional names. The probability that a web page will contain two names corresponding to two different individuals is quite low (~ca 7x10-8). Consequently co-present names form indicators of an individual's identity. These give accurate seed points, which are critical to the success of many clustering algorithms such as k-means (Jain et al. 1999)

1.3 Lexical Chain Text Similarity

Not all WePS pages contain multiple names, or even content in any form. Consequently we need to distinguish between pages that are similar in meaning to a page already in a seed cluster, those that refer to separate entities, and those to be discarded

This was done by comparing the conceptual similarity of the WePS pages using Roget's thesaurus as the conceptual inventory. The approach was described in Ellman (2000), where lexical chains are identified from each document using Roget's thesaurus. These chains are then unrolled to yield an attribute value vector of concepts where the values are given by repetition, type of thesaural relation found, and textual cohesion. Thus, we are not simply indexing by thesaural categories.

Vectors corresponding to different documents can be compared to give a measure of conceptual similarity. Roget's thesaurus typically contains one thousand sense entries divided by part of speech usage, giving a total of 6400 entries. Such vectors may be compared using many algorithms, although a nearest neighbor algorithm was implemented in Ellman (2000).

1.4 One Sense Per Discourse

UNN-WePS was based on a deliberate strategy that the success of an active disambiguation method needed to exceed its overall error rate in order to improve baseline performance. As such, simple methods that improved overall success modestly were preferred to complex ones that did not. Consequently, to reduce the search space, we used the 'one sense per discourse' heuristic (Gale et al. 1992). This assumes that one web page would not refer to two different individuals that share a name.

2 System Description

UNN-WePS was made up of three components, comprising modules to:

1. Create seed clusters that associated files with person names other than those being searched for.
2. Match similarity of unallocated documents to micro clusters using lexical chains derived from Roget's thesaurus.
3. Identify entities using single link agglomerative clustering algorithm.

In detail, a part of speech tagger (Coburn et al. 2007) was used to identify sequences of proper nouns. Person names were identified from these sequences using the following simple names 'grammar' coupled with data from the US Census (1990).

Name = [Title*][Initials | 1st name]+[2nd name]+

Figure 1: Regular Expression Name Syntax

We also used a gazetteer to forms seed clusters using data from the World Gazetteer (2007). This did not form part of the submitted system.

In the second step, conceptual similarity was determined using the method and tool described in Ellman (2000). Documents not allocated to seed clusters, were compared for conceptual similarity to all other documents. If similar to a document in a seed cluster, the unallocated document was inserted into the seed cluster. If neither document nor one to which it was similar too were in a seed cluster, they were formed into a new seed cluster. Finally if document has 'meaningful' content, but is not conceptually similar to any other it is stored in a singleton seed cluster otherwise, it is discarded.

In the final step, seed clusters were sorted by size and merged using a single link hierarchical agglomerative clustering algorithm to identify entities (Jain et al. 1999). The use of a single link means that a document can only be associated with one entity, which conforms to the 'one sense per discourse' heuristic.

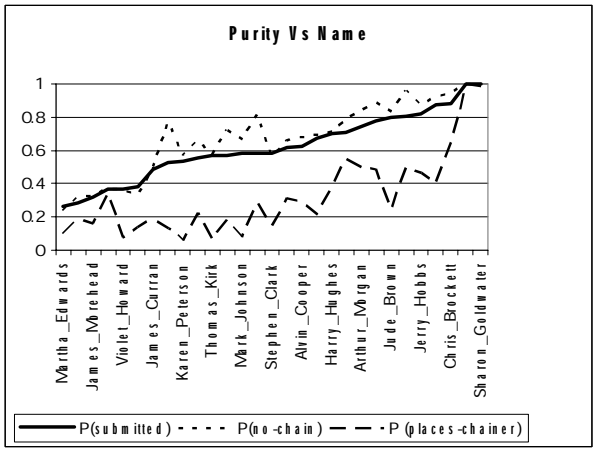
Further details of the UNN-WePS algorithm are given in figure 2 below.

```
FOREACH Person_Name
 1. TAG raw html Files with Part of Speech.
 2. IDENTIFY Generic Document Profiles using
      lexical chains in html Files.
 3. CONSTRUCT table T to associate person
      names with Files.
    a.  FOREACH File in Person_Name
          i. IDENTIFY Names in File
          ii. FOREACH Name in Names
                IF Name ≠ Person_Name
                STORE Name, File in T
 4. CREATE Seed clusters by inverting T to
      give files that are associated by co-
      present names
 5. MATCH Similarity of unallocated docu-
      ments to seed clusters
      a. FOREACH unallocated document D
         IF similar to a document in cluster C
           INSERT D into C
         ELSE IF similar to a non-clustered
         document D'
         CREATE D, D' as new cluster C'
         ELSE IF CONTAINS D > 200 words
              CREATE D as new cluster C''
              ELSE DISCARD D
 6. IDENTIFY entities using single link ag-
      glomerative clustering algorithm over
      seed clusters.
```

Figure 2: UNN-WePS Algorithm

3 Results

UNN-WePS achieved an average purity of 0.6, and inverse purity of 0.73 in Semeval Task 13, achieving seventh position out of sixteen competing systems (Artiles et al. 2007). However there was considerable variance in UNN-WePS results as shown in graph 1 below.



Graph 1: UNN-WePS purity performance

Graph 1 shows the purity scores for UNN-WePS on the Semeval 13 test data on three conditions: (1) as submitted (solid line), (2) using the gazetteer (dashed line), and (3) without the lexical chain based similarity matching (dotted line).

Note although the purity is lower when similarity matching is included the number of discarded documents is approximately halved.

An examination of the data suggests that where performance was especially poor it was due to genealogical data. Firstly this contains multiple individuals sharing the same name violating the 'one sense per discourse' heuristic. Secondly genealogical data includes birth and death information which was outside the scope of UNN-WePS. Furthermore, the large number of names confounds the statistical utility of co-present names.

4 Conclusion and Future Work

We have described a system, UNN-WePS that disambiguates individuals in web pages as required for Semeval task 13 (Artiles et al. 2007).

UNN-WePS was composed of three modules. The first formed seed clusters based on names present in web pages other than the individual for whom we are searching. The second used a lexical

chain based similarity measure to associates remaining files with clusters, whilst the third joined the clusters to identify identities using a single link hierarchical algorithm.

UNN-WePS performed surprisingly well considering the simplicity of its basic seeding algorithm. The use however of the 'one sense per discourse' heuristic was flawed. Names do re-occur across generations in families.

Genealogy is a popular Internet pastime, and web pages containing genealogy data frequently refer to multiple individuals that share a name at different time periods. As UNN-WePS did not account for time, this could not be detected. Furthermore, the large number of names in on-line genealogical data does lead to spurious associations.

As WePS was time limited, several extensions and refinements were envisaged, but not executed. Firstly, as described, the world gazetteer (2007) did not lead to performance improvements. We speculate therefore the disambiguation effect from using place names was exceeded by the ambiguity introduced by using them blindly. We note especially the inference between unidentified names (or street names, or building names) being interpreted as place data.

A further system deficiency was the lack of recognition of date data. This is essential to differentiate between identically named individuals in genealogical data.

Finally, we are currently experimenting with different clustering algorithms using the CLUTO toolkit (Karypis 2002) to improve on UNN-WePS baseline performance.

References

Adamic L.A. and Huberman B.A., 2002 *Zipf's law and the Internet*, *Glottometrics* 3, 2002, 143-150

Artiles, J., Gonzalo, J. and Sekine, S. (2007). *The SemEval-2007 WePS Evaluation: Establishing a benchmark for the Web People Search Task*. In Proceedings of Semeval 2007, Association for Computational Linguistics.

Ash, Russell 2006 *The top 10 of Everything* Hamlyn, Palazzo Bath UK

Bollegala, Danushka, Matsuo Yutaka Ishizuka Mitsuru *Disambiguating Personal Names on the Web using Automatically Extracted Key Phrases* Proc. ECAI 2006, pp.553-557, Trento, Italy (2006.8)

Coburn A, Ceglowski M, and Cuadrado J 2007 *Lingua::EN::Tagger, a Perl part-of-speech tagger for English text*. http://search.cpan.org/~acoburn/Lingua-EN-Tagger-0.13/

Ellman, Jeremy. 2000 *Using Roget's Thesaurus to Determine the Similarity of Texts*. PhD thesis, University of Sunderland [Available at http://citeseer.ist.psu.edu/ellman00using.html]

Gale, W., Church, K., and Yarowsky, D. (1992). *One sense per discourse*. In Proceedings of the Fourth DARPA Speech and Natural Language Workshop, pages 233--237.

Guha R. & Garg A. *Disambiguating People in Search*. Stanford University, 2004

Jain, A. K., Murty, M. N., and Flynn, P. J. 1999. *Data clustering: a review*. ACM Comput. Surv. 31, 3 (Sep. 1999), 264-323

Karypis G. 2002. *CLUTO: A clustering toolkit*. Technical Report 02-017, University of Minnesota. Available at: http://wwwusers.cs.umn.edu/~karypis/cluto/.

Leidner, Jochen L. (2004). *Toponym Resolution in Text: "Which Sheffield is it?"* in proc. 27th Annual International ACM SIGIR Conference (SIGIR 2004), Sheffield, UK.

Navigli, Roberto 2006. *Meaningful clustering of senses helps boost word sense disambiguation performance*. In *Proc. ACL* (Sydney, Australia, July 17 - 18, 2006).

Salton and Buckley 1990 *Improving Retrieval Performance by Relevance Feedback* JASIS 41(4) pp288-297

US Census 1990 http://www.census.gov/genealogy/names/names_files.html accessed 17th April 2007

Wan, X., Gao, J., Li, M., and Ding, B. 2005. *Person resolution in person search results: WebHawk*. in Proc. CIKM '05. ACM Press, New York, NY

World Gazetteer 2007 http://world-gazetteer.com/ accessed 17th April 2007