Domain-Independent Data Extraction: Person Names

Carl Christensen and Deryle Lonsdale
Brigham Young University

Background

Many problems stand in the way of true web annotation. A lack of standards in web layout, documentation, and format all contribute to the chaos that is the modern internet. But with many modern data extraction methods and annotation standards, much of the noisy data disappears. Software built for domain-specific data extraction has greatly simplified the problem of web annotation in such domains as obituaries, car ads, etc. Domain-independent data extraction, on the other hand, remains an open question. Existing ontology-based frameworks effectively extract data on a given semantic domain. Person names present a unique challenge because personal information is rarely domain-specific. Information about people clouds the internet, complicated by the wide variety of ways in which a person's name and information are referenced.

The WePS (Web Person Search) competition has been established to address the problem of extracting personal information—to build application that handle these data, and to determine the best approach to extracting a broad variety of data on any given person. The WePS website (http://nlp.uned.es/weps/) outlines the problem in retrieving personal information: "The user is […] forced either to add terms to the query (probably losing recall and focusing on one single aspect of the person), or to browse every document in order to filter the information about the person he/she is actually looking for."

The Data Extraction Group (DEG) at BYU has been engaged in this kind of work and research for over ten years (www.deg.byu.edu). In that time, the group has worked on developing conceptual-model-based ontologies and related applications[1]. These ontologies leverage keywords and phrases in building up a useful data structure that can be queried and used in extracting semantic information about a webpage. DEG efforts have combined many tools and application strategies to facilitate the management of different data formats and structures. In this paper, we discuss the practicality of the DEG approach, its strengths and shortcomings in domain-independent data extraction, and future work to better enable attribute extraction via such ontology based applications.

During 2007 the first iteration of the WePS competition was held. Sixteen systems participated in the competition which focused on clustering person's names into groups of web pages referring to the same individual. This year's competition included another related task—extracting attributes associated with person names. The attribute extraction (AE) task required the system to extract 18 attribute-value pairs ranging from the person's name(s) to their nationality, fax number, affiliations, relatives' names, and so on. It WePS participants received a new training corpus on which to develop or train the competing systems. The training corpus for the attribute subtask consisted of 17 search names and approximately 100 web pages per name. The corpus was human annotated for target attribute-value pairs that the systems would be expected to extract from each page, and included an evaluation script on which a system's precision, recall, and f-measure could be calculated. When the development phase which lasted October and November

was over, the competitors received a test set of unseen web pages and person names that the systems had to process. The competition organizers then conducted the final evaluation of the systems with another such gold-standard corpus.

As mentioned, the Data Extraction Group has conducted similar research for years. DEG software includes Ontos, OntologyEditor, OSMX, and many other useful tools which perform various ontology-based extraction tasks. Ontologies organize data into concepts and relationships and thus can present labels for words on a given semantic domain[2]. The structure of an ontology serves to organize data into searchable, indexable output. The DEG Ontos system takes a user-specified ontology and scans documents for values matching the labels and semantic categories referenced in the ontology. The bulk of DEG research hinges on the ability of software to leverage ontologies for accurate data extraction. These applications have been used for projects very similar to the WePS competition and indeed are well suited for this kind of task because of its ability to constrain and organize ontologies in a way that allows for both domain-specific and domain-independent data extraction.

Procedure

Our intent in participating in the WePS competition was to see how well the DEG system would perform in this type of system evaluation scenario. Since there was little time for system development, the goal was to employ the system in as close to an off-the-shelf manner as possible. In this section we sketch how the Ontos Semantic Annotator system was used.

The system takes as input an ontology and an HTML web page. After scrubbing the web page to remove extralinguistic content, the software searches the page using the specified ontology, locating and annotating the desired attributes. All attribute-value matches were output to an ASCII file formatted to match the specifications of the WePS evaluation scripts. Throughout the training/development period we ran the system on the training corpus, analyzed the results, and modified the system as necessary to improve system performance.

The ontology constraints allowed in the Ontos system provide a method of structuring and categorizing text based on lexical clues. The semantics of the given attributes are mapped out via knowledge sources, regular expressions, keywords, etc. These constraints fit well into the model of an extraction ontology—so termed because of the ontology's use to extract data. The search for eighteen attributes on varying semantic domains, all associated with a target person, lends itself to construction of an ontology bounded by the real-world constraints on personal information. As we built the ontology for this task, we identified the attributes as object sets with relationships to the main object set—the person in question. Thus values are extracted under the specified attribute and be linked with the person name building a relationship between the attributes and the main objet set. While much functionality that exists with the DEG Ontos system—and with extraction ontologies in general—is not implemented in this ontology, the structure and flexibility of this ontology enabled us to establish a robust framework for domain-independent data-extraction.

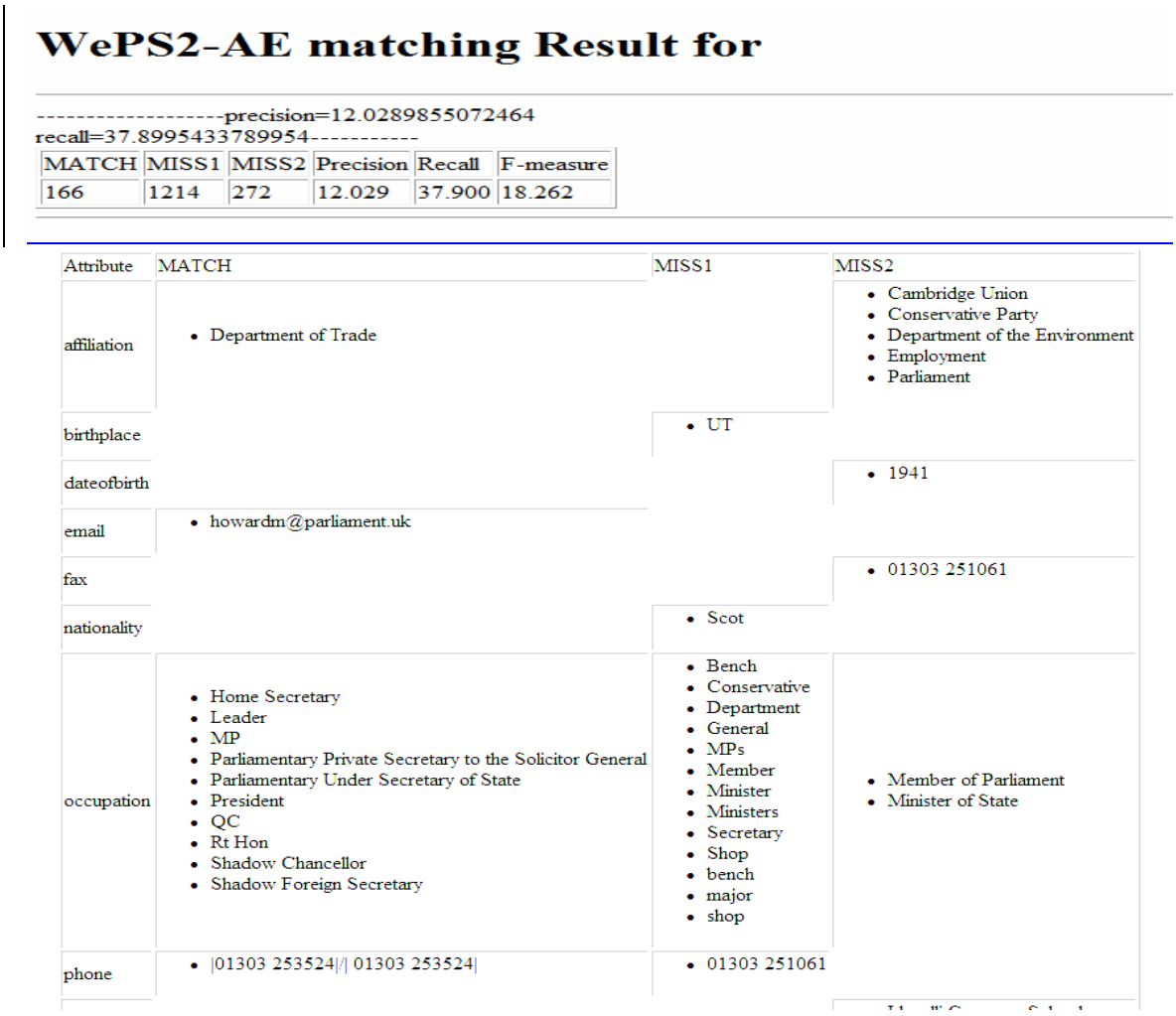
In order to utilize the DEG Ontos system in the WePS competition, we gathered and built knowledge source files as a reference for some of the attributes. The ontology-

based approach makes use of knowledge files, comparing the text of the target page against a repository of words or phrases stored in a data file. Much of our work in this effort involved populating the data repository with more complete knowledge for the required attribute values. For example, in building the files specifying possible values for the occupation and school attributes, we made use of publicly accessible data bases—all internet based. We gathered data from various online databases and resources including the U.S. Census Bureau, Wikipedia, livejounal.com, and various informational and commercial websites. After gathering lists from these sites, we formatted the data as necessary. The result was an enumeration of all possible values for the desired attributes—such as college degree or nationality. For other attributes, the data we collected represented only a small but strategically-chosen sampling of the possible values for the attribute. The dictionary file for schools listed over 70,000 institutions, which is still a small subset of possible values for the school and affiliation attributes. With the large amounts of data collected, and the various forms of formatting, erroneous words and bad data inevitably slipped into the files as well.

Results

During the training phase evaluation of the system's performance was possible via scripts and other tools made available to the participants by the organizers. For example, the extractions for each file for a given target person were easily visualized via an HTML report page like the one shown below. The figure below shows output after the evaluation scripts were run one text file output by Ontos. Here, all the dictionary files had been included. The results below are for all the web pages for one search person. The 37.9% recall shown below is calculated by taking the 166 matches found and dividing by the total number of possible matches. Thus the 166 matches found divided by the total of the matches plus the 272 missed matches (miss2) gives the recall percentage.. The precision is calculated by taking the number of matches divided by the total number of proposed matches. So 166 correct matches plus 1214 incorrect matches (miss1) severely dropped the precision to 12%. F-measure is a calculation based on a formula that combines the recall and precision scores for an overall rating.

Our experience involved negotiating the fine line between precision and recall—in particular with the dictionary files. The larger the data files, the more complete the recall. On the other hand, the larger the file, the more overlap and noise words enter the lexicon that can lead to increased recall but lower precision. For example, the first pass Ontos made on the training data, only a few lexicon files were included. These first passes resulted in higher precision, but low recall. Both measures were quite low with the average precision at 2.7% and the average recall at 3.5%. After many of the lexicon files were included the recall doubled to 8.6%, but the precision initially decreased to 1.7%. Many of the erroneous words accounted for the severe drop in precision, but many of the values matched with attributes were feasible if taken out of context. Other features of the Ontos system were then implemented to constrain the dictionary files matches and thus disallow much of the erroneous output.



WePS2-AE matching Result for

----------precision=12.0289855072464
recall=37.8995433789954----------

MATCH	MISS1	MISS2	Precision	Recall	F-measure
166	1214	272	12.029	37.900	18.262

Attribute	MATCH	MISS1	MISS2
affiliation	• Department of Trade		• Cambridge Union • Conservative Party • Department of the Environment • Employment • Parliament
birthplace		• UT	
dateofbirth			• 1941
email	• howardm@parliament.uk		
fax			• 01303 251061
nationality		• Scot	
occupation	• Home Secretary • Leader • MP • Parliamentary Private Secretary to the Solicitor General • Parliamentary Under Secretary of State • President • QC • Rt Hon • Shadow Chancellor • Shadow Foreign Secretary	• Bench • Conservative • Department • General • MPs • Member • Minister • Ministers • Secretary • Shop • bench • major • shop	• Member of Parliament • Minister of State
phone	• |01303 253524|/| 01303 253524|	• 01303 251061	

The system's performance improved drastically throughout the time spent improving the Ontos system. Through running the training data, we found many areas where the system's constraints could be utilized to boost precision and incorporate regular expressions to increase recall. After working out many the challenges associated with the large data set and domain-independent data extraction, our results increased substantially. Our final comprehensive run of the system over all the web pages resulted in an average of about 27%, with recall for some of the search names around 40%. Precision averaged much lower at 6.7%, with the best search results around 10%. Much of the precision problems stemmed from another serious problem of domain-independent data extraction. The DOM parser used in the system could not identify unofficial HTML tags present in many of the training data set's pages. When the system encountered these files, it failed to return any output. Once again, the gold standard pages and the evaluation scripts served to identify the problem.

Overlapping matches also presented a problem. This problem was largely unique to domain-independent, large scale data extraction. Initially, a feature in Ontos eliminated the longer of the two matches if two matches were found. This arbitrary rule demonstrates the scarcity of such a problem in domain specific data extraction. In domain-independent data extraction, on the other hand, such problems occur frequently.

For example, if ‘University of Arizona’ was a possible attribute value for School, and ‘Arizona’ was a possible attribute value for Birthplace, the system preferred the shorter match. Despite the initial problem, this evidenced Ontos’ ability to identify text that matched more than one attribute—a term could match both a regular expression and a lexicon file. After a simple alteration to the system, overlapping matches were evaluated separately for the different attributes, thus increasing recall.

Ontos’ cardinality constraints further controlled of attribute values. The values for birth date could thus be bounded by these constraints to some reasonable number. The constraints are enforced on the relationship sets to constrain the output. Thus the constraints on the WePS Person—the main object set—control the number of relationships that are drawn between the person and the various attributes. Constraints imposed on the attribute value pairs represent the possible number of values that are assigned or found for each attribute. Thus a constraint of 1:* signifies the idea that at least one attribute value will be found and that there is no bound on the total number of attribute values that can be found. Notation of optional participation, 0:*, indicates any number of values are extracted, or none at all. For example, while a person can have any number of mentors or affiliations, a constraint is set on birth date to both allow differing information while bounding the software from picking up any number of dates as birth date values. While in practicality, only one birth date exists for each person, such a cardinality constraint disallows variance and a margin for error between pages. In the WePS task, we discovered that the constraints of the Ontos’ system were not fully functional—another sign that domain-dependent, small data sets do not require the same functionalities as domain-independent. Ideally, the ontology system should make use of these restraints as an object set is established for each person name, allowing the given number of instances of the value across pages.

Keyword constraints, or required context, give the system the required lexical clues to help identify terms and values in text. The attribute value for the various kinds of information thus relies on plausible lexical identifiers to finally categorize a term in the ontology. We specified word boundaries and required words to further refine the extraction process. For example, a number or text that matched a given regular expression was checked against the keyword constraints. If the expression was within *n* words of the keyword ‘birthday’ or ‘birth date’ or ‘born,’ it was assumed that the date referenced the birth date. This lexical context is crucial to domain-independent data extraction as it allows the lexicon files to be populated with overlapping data that can then be eliminated by checking any number of plausible context clues before outputting results.

Text formatting served as an obstacle to correct extraction. Much of the existing work in data extraction focuses on the formatting of tables or markup tags. Domain-independent data extraction brings with it the added difficulty of free-form text extraction. Where the Ontos system had previously been set up to view paragraph breaks as new records, or the end of a main object set in the extraction ontology, the sample web pages consisted of varying types of document structures. Some pages consisted of a table of information about the search person, while others were essay or encyclopedia text in a plain text format. We accordingly allowed multiple paragraphs in a webpage to build the same record and relied on our context constraints to validate that extracted information was in reference to the person name in question (i.e. that a reference to the University of

Arizona was about the affiliation or school of the search person instead of another person mentioned in the webpage).

We had no previous values to compare our results to, since no similar objective attribute extraction evaluation had apparently ever been carried out on a task like the WePS person task. Still, due to the difficulties we had encountered along the way, and the fairly low numbers we were achieving on the training data, we decided not to submit the final results to the WePS organizers for the final tally. Results of those who did submit their final test runs have not yet been made public.

Related work and future work

Other work on domain-independent data extraction proposes alternate methods which also attempt to constrain extraction tools. Researchers from the P.E.T research centre at Mandya highlight the ability of ontologies to organize and extract data from unstructured text documents [3]. Researchers at the Database and Artificial Intelligence Group at Vienna University of Technology focus on the ability to use visual clues to facilitate plausible data extraction[4]. Many researchers are also looking into automatic and semi-automatic ontology generation.

Much future work exists for domain-independent data extraction, especially in the context of person names. After observing the results of the Ontos system, we have identified ways to improve the system's performance on domain-independent attribute extraction. The further development of the context constraints could also lead to dramatic improvement in the precision of such extraction ontologies. Work must be done to enable the system to take keywords and context words as a probability that a value or phrase matches a given attribute, and thus be taken as one of many considerations.

The field of data mining also overlaps largely with the field of data extraction, and the nature of this competition in particular. Researchers in data mining have also discussed the difficulty of identifying data in text form and more complex data forms[5]. Data mining also relies on the ability of the system to structure the contents of the material into categories through a rule based approach.

Future work in domain-independent data extraction also extends to person name disambiguation. The correct identification of these attribute values for person names makes the identification of individuals with the same name a much more manageable problem. A web page referring to a person with a different birth date and birthplace are assumed to be a different individual.

Conclusion

In this project, we implemented applications that have been developed by the Data Extraction Group in extracting attribute values for the WePS competition. We showed the plausibility of ontology based data extraction in domain-independent data extraction for person name information. The system demonstrated its robust nature, and its ability to allow lexical clues and constraints. The system also handled the searching and matching of large lexicon files without a noticeable effect on the time a webpage annotation takes. We also identified areas of future research and work to increase the recall and precision for any number or semantic category of given attributes.

References

1. Embley, D.W. et. al. (1999). Conceptual-Model-Based Data Extraction from Multiple-Record Web Pages. *Data & Knowledge Engineering*, 31, 227-251.

2. Cheng , C. K., Pan, X. S., & Kurfess, F. (2003). Ontology-based Semantic Classification of Unstructured Documents. *Lecture Notes in Computer Science* (pp. 441-458). Berlin: Springer.

3. Murali , H.L & Shashirekha1 S. (2007). Ontology Based Structured Representation for Domain Specific Unstructured Documents. *Proceedings of the International Conference on Computational Intelligence and Multimedia Applications*, 1, 50-54.

4. Gatterbauer, W., Bohunsky, P., Herzog, M., Krupl B., & Pollak, B. (2007). Towards Domain-Independent Information Extraction from Web Tables. *Proceedings of the 16th international conference on World Wide Web*, 71 – 80.

5. Yang, Q. & Wu, X.. (2006). 10 Challenging Problems in Data Mining Research. *International Journal of Information Technology & Decision Making*, 5, 597–604.