UA-ZSA: Web Page Clustering on the basis of Name Disambiguation

Zornitsa Kozareva, Sonia Vazquez, Andres Montoyo
DLSI, University of Alicante
Carretera de San Vicente S/N
Alicante, Spain
03080
zkozareva,svazquez,montoyo@dlsi.ua.es

Abstract

This paper presents an approach for web page clustering. The different underlying meanings of a name are discovered on the basis of the title of the web page, the body content, the common named entities across the documents and the sub-links. This information is feeded into a K-Means clustering algorithm which groups together the web pages that refer to the same individual.

1 Introduction

Ambiguity is the task of building up multiple alternative linguistic structures for a single input. Most of the approaches focus on word sense disambiguation (WSD), where the sense of a word has to be determined depending on the context in which it is used.

The same problem arises for named entities shared by different people or for grandsons named after their grandparents. For instance, querying the name "Michael Hammond" in the World Wide Web where there are huge quantities of massive and unstructured data, a search engine retrieves thousands of documents related to this name. However, there are several individuals sharing the name "Michael Hammon". One is a biology professor at the University of Arizona, another is at the University of Warwick, there is a mathematician from Toronto among others. The question is which one of these referents we are actually looking for and interested in. Presently, to be able to answer to this question, we have to skim the content of the documents and retrieve the correct answers on our own.

To automate this process, the named entities can be disambiguated and the different underlying meanings of the name can be found. On the basis of this information, the web pages can be clustered together and organized in a hierarchical structure which can ease the documents' browsing. This is also the objective of the Web People Search (WePS) task (Artiles et al., 2007). What makes the WePS task even more challenging is the fact that in contrast to WSD where the number of senses of a word are predefined, in WePS we do not know the exact number of different individuals.

For the resolution of the WePS task, we have developed a web page clustering approach using the title and the body content of the web pages. In addition, we group together the documents that share many location, person and organization names, as well as those that point out to the same sub-links.

The rest of the paper is organized as follows. In Section 2 we describe various approaches for name disambiguation and discrimination. Our approach is shown in Section 3, the obtained results and a discussion are provided in Section 4 and finally we conclude in Section 5.

2 Related Work

Early work in the field of name disambiguation is that of (Bagga and Baldwin, 1998) who proposed cross-document coreference resolution algorithm which uses vector space model to resolve the ambiguities between people sharing the same name. The approach is evaluated on 35 different mentions of John Smith and reaches 85% f-score.

Mann and Yarowski (2003) developed an unsu-

Proceedings of the 4th International Workshop on Semantic Evaluations (SemEval-2007), pages 338–341,
Prague, June 2007. ©2007 Association for Computational Linguistics

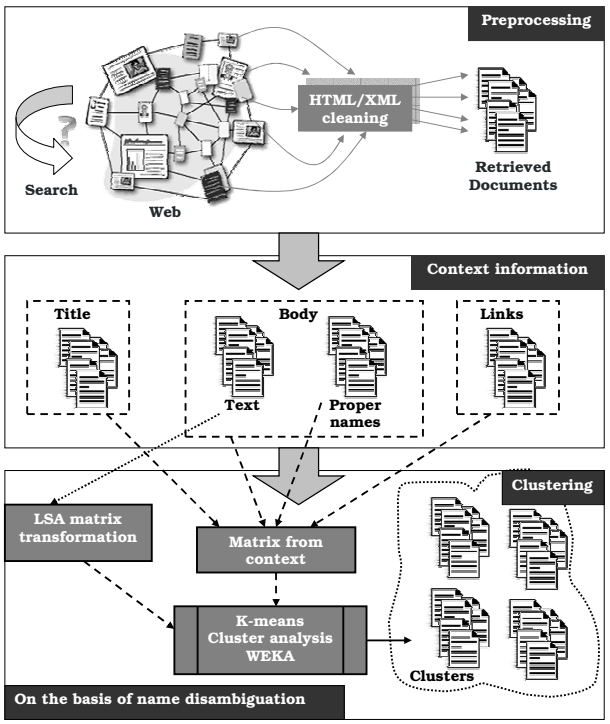

Figure 1: Architecture of the WePS System

pervised approach to name discrimination where biographical features (age, date of birth), familiar relationships (wife, son, daughter) and associations (country, company, organization) are considered. Therefore, in our approach we use person, organization and location names in order to construct a social similarity network between two documents.

Another unsupervised clustering technique for name discrimination of web pages is that of Pedersen and Kulkarni (2007). They used contextual vectors derived from bigrams, and measured the impact of several association measures. During the evaluation, some names were easily discriminable compared to others categories for which was even difficult to find and obtain discriminative feature. We worked with their unigram model (Purandare and Pedersen, 2004) to cluster the web pages using the text content between the title tags.

3 Web Person Disambiguation

Our web people clustering approach is presented in Figure 1 and consists of the following steps:

- HTML cleaning: all *html* tags are stripped away, the *javascript* code is eliminated, the non closed WePS tags are repaired, the missing begin/end body tags are included and then the content between the title, the body and the anchor tags is extracted.

- name matching: the location, person and organization names in the body texts are identified with the GATE[1] system (Cunningham, 2005). Each named entity of a document is matched with its corresponding named entity category from the rest of the web pages. This information is used to calculate the social semantic similarity of the person, the location and the organization names. Our hypothesis is that documents with similar names tend to refer to the same individual. The output of this module is a matrix with binary values, where 1 stands for the documents which share more than the half of their proper names, and 0 otherwise.

- links: for each document, we extract the links situated between the anchor tags. Since the links are too specific, we wrote an url function which transform a given web page d_1 with URL http://www.cs.ualberta.ca/~lindek/index.htm into www.cs.ualberta.ca/~lindek, and the web page d_2 with URL http://www.cs.ualberta.ca/~lindek/demos.htm into www.cs.ualberta.ca/~lindek. According to our approach, the two web pages d_1 and d_2 are linked to each other if their link structures (LS) intersect, that is $LS(d1) \bigcap LS(d2) \neq 0$. The output of this module is a matrix with binary values, where 1 stands for two web pages having more than 3 links in common and 0 otherwise.

- titles: for each document, we extract the text between the title tags. We create a unigram matrix which is feed into SenseClusters[2]. We use automatic cluster stopping criteria with the gap statistics which groups the web pages into several clusters according to the context of the titles. From the obtained clusters, we generate a new matrix with binary values, where 1 corresponds to the documents which were put in the

[1] http://sourceforge.net/projects/gate
[2] http://marimba.d.umn.edu/cgi-bin/SC-cgi/index.cgi

same cluster according to SenseClusters and 0 otherwise.

- bodies: the text between the body tags is extracted, tokenized and the part-of-speech (POS) tags [3] are determined. The original text is transformed by encoding the POS tag information as follows: "*water#v the#det flowers#n and#conj pass#v me#pron the#det glass#n of#prep water#n*". This corpus transformation is done, because we want the Latent Semantic Analysis (LSA) module to consider the syntactic categories of the words and to construct a more reliable semantic space. For instance, in the example above, there are two different representations of *water*: the noun and the verb, while without the corpus transformation LSA sees only the string *water*.

- LSA[4]: the semantic similarity score for the web-pages is calculated with Latent Semantic Analysis (LSA). From the encoded body texts, we build up a matrix, where the rows represent the words of the web-page collection, the columns stand for the web-pages we want to cluster and the cells show the number of times a word of the corpus occurs in a web page. In order to reduce the noise and the data sparsity, we apply the Singular Value Decomposition algorithm by reducing the original vector space into 300 dimensions. The output of the LSA module is a matrix, which represents the semantic similarity among the web pages.

- knowledge combination: the outputs of the name matching, link, title and body modules are combined into a new matrix 100×400 dimensional matrix. The rows correspond to the number of web pages and the columns represent the obtained values of the link, title, body and name modules. This matrix is fed into the K-means clustering algorithm which determines the final web page clustering.

- K-means[5]: the clustering of N web pages into K disjoint subsets S_j containing N_j data points is done by the minimization of the sum-of-squares criterion $J = \sum_{j=1}^{K} \sum_{n \in S_j} |x_n - mu_j|^2$, where x_n is a vector representing the nth data point and mu_j is the geometric centroid of the data points in S_j. The information matrix from which the web page clustering is performed includes the similarity information for the title, link, proper name and body. The current implementation of K-means (Witten and Frank, 2005) does not have an automatic cluster stopping criteria, therefore the number of clusters is set up manually.

[3] http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/
[4] infomap-nlp.sourceforge.net/
[5] http://www.cs.waikato.ac.nz/ml/weka/

4 Results and Discussion

Table 1 shows the obtained results for the test data set. The average performance of our system is 56% and we ranked on 10-th position from 16 participating teams. Although, we have used different sources of information and various approximations, in the future we have to surmount a number of obstacles.

One of the limitations comes from the usage of the text snippets situated between the body tags. There are a number of web pages which do not contain any text. The semantic space for these documents cannot be built with LSA and their similarity score is zero.

Despite the fact that we have eliminated the stop words from the documents and we have transformed the web pages by encoding the syntactic categories, the classification power of LSA was different for the ambiguous names and for the web pages. To some extend this is due to the varying number of words in the web pages. In the future, we want to conduct experiments with a fixed context windows for all documents.

In this task, the number of senses (e.g. number of different individuals that share the same name) is unknown, and one of the major drawbacks in our approach is related to the setting up of the number of clusters. The K-Means clustering algorithm we used, did not include an automatic cluster stopping criteria, and we had to set up the number of clusters manually. To be able to do that, we have observed the average number of clusters per name in the trial data. We have evaluated the performance of our approach with several different numbers of clusters. According to the obtained results, the best clusters are 25 and 50. We used the same number

Name	Purity	Inverse Purity	F α=0.5	F α=0.2
Mark Johnson	0,55	0,74	0,63	0,69
Sharon Goldwater	0,96	0,23	0,37	0,27
Robert Moore	0,36	0,67	0,47	0,57
Leon Barrett	0,62	0,51	0,56	0,52
Dekang Lin	0,99	0,43	0,60	0,49
Stephen Clark	0,52	0,75	0,62	0,69
Frank Keller	0,38	0,67	0,48	0,58
Jerry Hobbs	0,54	0,63	0,58	0,61
James Curran	0,53	0,61	0,57	0,59
Chris Brockett	0,73	0,40	0,51	0,44
Thomas Fraser	0,66	0,57	0,61	0,58
John Nelson	0,68	0,76	0,72	0,74
James Hamilton	0,56	0,60	0,58	0,59
William Dickson	0,59	0,78	0,67	0,73
James Morehead	0,36	0,64	0,46	0,56
Patrick Killen	0,56	0,69	0,62	0,66
George Foster	0,46	0,70	0,56	0,64
James Davidson	0,58	0,71	0,64	0,68
Arthur Morgan	0,77	0,47	0,59	0,51
Thomas Kirk	0,26	0,90	0,41	0,60
Patrick Killen	0,56	0,69	0,62	0,66
Harry Hughes	0,66	0,54	0,59	0,56
Jude Brown	0,64	0,63	0,64	0,63
Stephan Johnson	0,56	0,80	0,66	0,73
Marcy Jackson	0,40	0,73	0,52	0,63
Karen Peterson	0,56	0,72	0,63	0,68
Neil Clark	0,68	0,36	0,47	0,40
Jonathan Brooks	0,53	0,76	0,63	0,70
Violet Howard	0,58	0,75	0,65	0,71
Global average	0,58	0,64	0,58	0,60

Table 1: Evaluation results

of clusters for the test data, however this is a rough parameter estimation.

5 Conclusion

Person name disambiguation is a very important task whose resolution can improve the performance of the search engine by grouping together web pages which refer to different individuals that share the same name.

For our participation in the WePS task, we presented a name disambiguation approach which uses only the information extracted from the web pages. We conducted an experimental study with the trail data set, according to which the combination of the title, the body, the proper names and sub-links reaches the best performance. Our current approach can be improved with the incorporation of automatic cluster stopping criteria.

So far we did not take advantage of the document ranking and the returned snippets, but we want to incorporate this information by measuring the snippet similarity on the basis of relevant domain information (Kozareva et al., 2007).

Acknowledgements

Many thanks to Ted Pedersen for useful comments and suggestions. This work was partially funded by the European Union under the project QALLME number FP6 IST-033860 and by the Spanish Ministry of Science and Technology under the project TEX-MESS number TIN2006-15265-C06-01.

References

J. Artiles, J. Gonzalo, and S. Sekine. 2007. The semeval-2007 weps evaluation: Establishing a benchmark for the web people search task. In *Proceedings of Semeval 2007, Association for Computational Linguistics*.

A. Bagga and B. Baldwin. 1998. Entity-based cross-document coreferencing using the vector space model. In *Proceedings of ACL*, pages 79–85.

H. Cunningham. 2005. Information Extraction, Automatic. *Encyclopedia of Language and Linguistics, 2nd Edition*.

Z. Kozareva, S. Vazquez, and A. Montoyo. 2007. The usefulness of conceptual representation for the identification of semantic variability expressions. In *Proceedings of the Eighth International Conference on Intelligent Text Processing and Computational Linguistics, (CICLing-2007)*.

G. Mann and D. Yarowsky. 2003. Unsupervised personal name disambiguation. In *Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003*, pages 33–40.

T. Pedersen and A. Kulkarni. 2007. Discovering identities in web contexts with unsupervised clustering. In *Proceedings of the IJCAI-2007 Workshop on Analytics for Noisy Unstructured Text Data*.

A. Purandare and T. Pedersen. 2004. Senseclusters - finding clusters that represent word senses. In *AAAI*, pages 1030–1031.

I. Witten and E. Frank. 2005. *Data Mining: Practical machine learning tools and techniques*, volume 2. Morgan Kaufmann.