Dynamic Threshold for Clustering Person Names

Roberto Zanoli, Francesco Corcoglioniti and Christian Girardi

Fondazione Bruno Kessler,
38121 Trento, Italy
{zanoli,corcoglio,cgirardi}@fbk.eu

Abstract. Nowadays, surfing the Internet and looking for persons seems to be one of the most common activities of Internet users. However person names could be highly ambiguous and consequently search results are often a collection of documents about different people sharing the same name. In this paper a system able to identify person names in different documents which refer to the same person entity is presented. Differently from other systems which adopt a fixed similarity threshold to group documents talking about the same person, the presented approach uses a threshold capable of changing its value on the basis of the ambiguity of the name as estimated by using external resources (i.e. phonebooks). For each name the algorithm was provided with a specific threshold value and with a rich set of features (e.g. Named Entities) extracted from the document where the person name is mentioned; the performance of the system was tested taking part in the News People Search (NePS) task at Evalita 2011.

Keywords: Information extraction, Clustering, Person Name Disambiguation, Cross-Document Coreference

1 Introduction

According to a study of the query log of Altavista [5], around 17% of the queries contain personal names, and surfing the Internet and looking for persons seems to be one of the most common activities of Internet users. However person names could be highly ambiguous and consequently search results are often a collection of documents about different people sharing the same name. Hence, the task of disambiguating personal names across documents plays an important role.

Cross-document coreference resolution is the task of identifying person names in different documents which refer to the same person entity, and tasks such as Web People Search evaluation exercise (WePS)[1] and News People Search (NePS)[2] tasks aim at evaluating systems for cross-document coreference.

A study of [1] shows that the Hierarchical Agglomerative Clustering (HAC) was the most popular algorithm used in WePS, with the output number of

[1] http://nlp.uned.es/weps/weps-2
[2] http://www.evalita.it/2011/tasks/NePS

clusters (i.e. the different number of persons) determined by a fixed similarity threshold; the threshold determines how close two elements (i.e. documents or clusters) have to be so as to be grouped together. They set a threshold and stop clustering once the distance between elements is above the threshold.

In contrast to such approach, this paper describes a system with a similarity threshold which is not fixed for all the names, but depends on the ambiguity of the name as estimated by using external resources (i.e. phonebooks); the threshold value increases for not ambiguous names whereas it decreases for names which are ambiguous. As the clustering algorithm is concerned, the QT [4] was used. It is an alternative method of partitioning data, invented for gene clustering. It requires more computing power than k-means and the setting of the cluster diameter (the maximal distance between any two elements of the cluster), but does not require specifying the number of clusters a priori. The system is also able to exploit a rich set of features (e.g. Named Entities and keyphrases co-occurring in the same document of the name) whereas its performance was evaluated taking part in the News People Search task at Evalita 2011.

The remainder of the paper is organized as follows: Section 2 briefly describes the NePS task at Evalita 2011 where the presented system took part in, Section 3 explains the reason of having a dynamic threshold, Section 4 describes the QT algorithm and its implementation, Section 5 and Section 6 analyze in detail the experiments done and the results obtained by the system. Section 7 reports the conclusions.

2 NePS Task at Evalita 2011

NePS task requires the systems to cluster together documents extracted from Italian newspapers talking of the same person: for each person name, and given a set of documents in input, systems were asked to group the documents, so that each cluster only contains the documents which refer to the same person.

Participants were provided with a development set of 105 Group names so as to configure their systems whereas a test data of 103 Group names was used to evaluate them. Here a Group Name is thought to be a complete name, i.e. a pair First-Name Last-Name (e.g. Paolo Rossi, Diego Armando Maradona) which can be shared by different persons. Both the development set and the test set are structured along the two orthogonal variables of entity fame (according to annotators knowledge about the entities) and Group name ambiguity (the ambiguity of the name in the corpus).

NePS evaluation was then carried out using WePS-2 official scorer and metrics (Bcubed precision, recall, and F1 measure).

3 Dynamic Threshold

Frequently the output number of clusters of systems (e.g. most of the WePS systems) is decided by considering the documents where the name appears and a fixed similarity threshold (a unique value is used for all the Group names),

which determines how close two elements (i.e. documents or clusters) have to be so as to be grouped together. The difficulty of the task stems in part from its reliance on world knowledge. To exemplify, consider the following text fragment where the contexts which contain *Luca Cordero di Montezemolo* are different and where the decision to cluster them together is difficult.

- Luca Cordero di Montezemolo's career began at the wheel of a Fiat 500.

- On 27 May 2004, Luca Cordero di Montezemolo became president of Confindustria.

However to a human, having the world knowledge that the ex-president of Confindustria *Luca Cordero di Montezemolo* was also a driver would be helpful for establishing the coreference relation between the two names; such knowledge may be also limited to the fact that names have different degrees of ambiguity (e.g. *Luca Cordero di Montezemolo* is not an Italian common name and therefore the documents might talk about the same person). Differently to *Luca Cordero di Montezemolo*, *Paolo Rossi* is an Italian common name. The chance that many different persons carry this name is high.

As reported by [3], PagineBianche[3] (i.e. the Italian phonebook) could be a good indicator of the ambiguity of a name in the NePS task: the more the ambiguity of a name in PagineBianche, the more the ambiguity of the name in NePS. Under this assumption a threshold decided on the basis of the ambiguity of a name determined by looking up the name in PagineBianche, has been studied. Paragraph 5 reports results when a 2-level threshold was applied.

4 Clustering Method

QT [4] is an alternative method of partitioning data, invented for gene clustering. In contrast to k-means, it is deterministic and does not require specifying the number of clusters a priori. On the other hand the algorithm requires the a priori specification of the diameter (the distance between each pair of elements) and that the minimum number of elements in each cluster has to be specified. Concerning its complexity, it depends on $O(n^3)$.

Algorithmic steps for QT clustering

- Initialize the diameter for clusters and the minimum cluster size.
- Build a candidate cluster for each element by including the closest one, the next closest, and so on, until the distance of the cluster surpasses the diameter.
- Save the candidate cluster which contains the highest number of elements as the first true cluster, and remove all elements in the cluster from further consideration.
- Repeat with the reduced set of elements until no more cluster can be formed having the minimum cluster size.

[3] www.paginebianche.it

5 Experiments

According to Bagga [2] each name to be clustered can be represented by using a vector of features extracted from the document where the name is extracted from. As far as this work is concerned, these features include:

- Topic of the Document: the topic the document is talking about (e.g. sport, gossip).
- Keyphrases: expressions, either single words or phrases, describing the most important concepts of a document (e.g. administrative committee, member of Parliament, reduction in tax).
- professional Category: professional category which is thought to belong to the name (e.g. president, journalist, football player).
- Named Entities: person, organization and location names (e.g. Bill Clinton, Ferrari, Rome).
- Ontology ID: approximate person name identier provided by automatically linking person names to a contextualized ontology describing 30493 persons relevant to the Trentino region and national-level as well.

Below an example of this representation for some occurrences of Kofi Annan, Mario Monti, Adriano Celentano is reported:

Table 1. Each of the name to be clustered was represented with a vector of features extracted from the document where it appears.

Name	Topic	Keyphrases	prof. Category	Named Entities	Ontology ID
Kofi Annan	politics	missile		Iraq	1345
Mario Monti	economy	ministers	premier	Rome	2656
Mario Monti	politics		commissioner	EU	2656
Adriano Celentano	tv	episode	singer	RAI	3768

The linking to the ontology was made by following the method of [6] whereas all the other features were produced by TextPro[4].

Different approaches to document representation may result from different choices as to how features weights should be computed. A common choice is Inverse Document Frequency (IDF) where one intuition is at play: the more documents a feature fk occurs in, the smaller its contribution is in characterizing the semantics of a document in which it occurs.

$$idf\left(fk\right) = \begin{cases} log\frac{|Tr|}{\#Tr(fk)} & if\#\left(fk\right) > 1 \\ 0 & \text{otherwise} \end{cases}$$

[4] textpro.fbk.eu

#Tr(fk) denotes the document frequency of feature fk, that is, the number of documents in Tr in which fk occurs. To make the weights fall in the [0,1] interval and for the documents to be represented by vectors of equal length, the weights resulting from idf were normalized by cosine normalization.

$$wkj = \frac{idf(fk,dj)}{\sqrt{\sum_{s=1}^{\|} T\|(idf(fs,dj))^2}}$$

The similarity between documents was then computed with the dot product of their respective vectors which corresponds to their cosine similarity (i.e. the cosine of the angle α that separates the two vectors).

Names were first categorized in 2 groups based on their ambiguity: names occurring less than 3 times in the phonebook were considered not ambiguous names, whereas those occurring more than 2 were considered ambiguous names. A threshold that maximizes the performance of the clustering algorithm on the development set was then used to cluster each of the 103 Group names of the test set: the diameter parameter of the QT algorithm was set to 0.98 for not ambiguous names, and to 0.90 for ambiguous ones. Eventually the minimum cluster size was assigned to 0 given that systems were asked to clusterize all the articles.

6 Results

Table 2 reports the performance of the system when it was tested on the NePS test set:

Table 2. Bcubed precision, recall, and F1 measure for different levels of ambiguity of a name: no ambiguity, medium ambiguity and high ambiguity.

	All			**no ambiguity**			**medium ambiguity**			**high ambiguity**		
	Pr	**Re**	**F1**	**Pr**	**Re**	**F1**	**Pr**	**Re**	**F1**	**Pr**	**Re**	**F1**
FBK	0.89	0.97	0.93	1.00	0.99	0.99	0.89	0.95	0.92	0.71	0.96	0.82
ALL-IN-ONE	0.84	1.00	0.91	1.00	1.00	1.00	0.86	1.00	0.93	0.56	1.00	0.72

ALL-IN-ONE means that all the provided documents for a specific Group name are taken to be related to the same person (i.e. only one cluster) and gives the highest possible Recall measure. Organizers reported results in terms of Bcubed precision, recall, and F1 measure and performances were distinguished considering the ambiguity of a name in the corpus: no ambiguity (there is only a person carrying that specific name), medium ambiguity (the name is shared by 2 or 3 different persons), high ambiguity (more than 3 persons have the same name). It should be noticed that even though the overall performance of the ALL-IN-ONE baseline lies close to the system one, differences exist when the 3 different levels of ambiguity of a name are taken into account. For names that are not so ambiguous (no ambiguity, and medium ambiguity in the table) the

system seems to be able to stays in the wake of ALL-IN-ONE whereas for names which are ambiguous the difference is more definite, 10 points of F1 in favor of the described system.

7 Conclusion

A system for coreference resolution and its participation in the Evalita 2011 evaluation campaign has been described. Each person name to be clustered was represented by using a vector of features extracted from the document where the name appears and a dynamic threshold depending on the ambiguity of the name was adopted. The threshold determines how close two documents have to be so as to be grouped together and it was estimated looking up the name in PagineBianche which is the Italian phonebook. The QT algorithm able to exploit these threshold values was then used to clusterize person names.

In spite of the characteristics of the NePS corpus rewarding ALL-IN-ONE, the proposed system appears to be good at emulating the high baseline values for names with a lower level of ambiguity while in contrast best results in favor of the system stand out for ambiguous names.

As to future work, a comparative study of different feature sets is going to be done whereas a measure to evaluate the statistical significance of the presented system against the ALL-IN-ONE baseline will be adopted.

Acknowledgments. This research was supported by the the LiveMemories project funded by the Provincia Autonoma of Trento.

References

1. Artiles, J., Gonzalo, J., Sekine, S.: WePS 2 Evaluation Campaign: overview of the Web People Search Clustering Task. In: 18th WWW Conference (2009)
2. Bagga, A., Baldwin, B.: Entity-Based Cross-Document Coreferencing Using the Vector Space Model. In: Thirty-Sixth Annual Meeting of the Association for Computational Linguistics and Seventeenth International Conference on Computational Linguistics (1998)
3. Bentivogli, L., Girardi, C., Pianta, E.: Creating a Gold Standard for Person Cross-Document Coreference Resolution in Italian News. In: LREC 2008 Workshop on Resources and Evaluation for Identity Matching, Entity Resolution and Entity Management (2008)
4. Heyer, L.J. and Kruglyak, S. and Yooseph, S.: Exploring Expression Data: Identification and Analysis of Coexpressed Genes. Genome Research. 9, 1106–1115 (1999)
5. Spink, A., Jansen, B., Pedersen, J.: Searching for people on web search engines. Journal of Documentation, 60, 266–278 (2004)
6. Tamilin, A., Magnini, B., Serafini, L.: Leveraging Entity Linking by Contextualized Background Knowledge: A case study for news domain in Italian. In: 6th Workshop on Semantic Web Applications and Perspectives (SWAP10), Bressanone, Italy (2010)