Experiments on Semantic-based Clustering for Cross-document Coreference

Horacio Saggion
Department of Computer Science
University of Sheffield
211 Portobello Street - Sheffield, England, UK, S1 4DP
Tel: +44-114-222-1947
Fax: +44-114-222-1810
saggion@dcs.shef.ac.uk

Abstract

We describe clustering experiments for cross-document coreference for the first Web People Search Evaluation. In our experiments we apply agglomerative clustering to group together documents potentially referring to the same individual. The algorithm is informed by the results of two different summarization strategies and an off-the-shelf named entity recognition component. We present different configurations of the system and show the potential of the applied techniques. We also present an analysis of the impact that semantic information and text summarization have in the clustering process.

1 Introduction

Finding information about people on huge text collections or on-line repositories on the Web is a common activity. In ad-hoc Internet retrieval, a request for documents/pages referring to a person name may return thousand of pages which although containing the name, do not refer to the same individual. Cross-document coreference is the task of deciding if two entity mentions in two sources refer to the same individual. Because person names are highly ambiguous (i.e., names are shared by many individuals), deciding if two documents returned by a search engine such as Google or Yahoo! refer to the same individual is a difficult problem.

Automatic techniques for solving this problem are required not only for better access to information but also in natural language processing applications such as multidocument summarization, question answering, and information extraction. Here, we concentrate on the Web People Search Task (Artiles et al., 2007) as defined in the SemEval 2007 Workshop: a search engine user types in a person name as a query. Instead of ranking web pages, an ideal system should organise search results in as many clusters as there are different people sharing the same name in the documents returned by the search engine. The input is, therefore, the results given by a web search engine using a person name as query. The output is a number of sets, each containing documents referring to the same individual. The task is related to the coreference resolution problem disregarding however the linking of mentions of the target entity inside each single document.

Similarly to (Bagga and Baldwin, 1998; Phan et al., 2006), we have addressed the task as a document clustering problem. We have implemented our own clustering algorithms but rely on available extraction and summarization technology to produce document representations used as input for the clustering procedure. We will shown that our techniques produce not only very good results but are also very competitive when compared with SemEval 2007 systems. We will also show that carefully selection of document representation is of paramount importance to achieve good performance. Our system has a similar level of performance as the best system in the recent SemEval 2007 evaluation framework. This paper extends our previous work on this task (Saggion, 2007).

2 Evaluation Framework

The SemEval evaluation has prepared two sets of data to investigate the cross-document coreference problem: one for development and one for testing. The data consists of around 100 Web files per person name, which have been frozen and so, can be used as an static corpus. Each file in the corpus is associated with an integer number which indicates the rank at which the particular page was retrieved by the search engine. In addition to the files themselves, the following information was available: the page title, the url, and the snippet. In addition to the data itself, human assessments are provided which are used for evaluating the output of the automatic systems. The assessment for each person name is a file which contains a number of sets where each set is assumed to contain all (and only those) pages that refer to one individual. The development data is a selection of person names from different sources such as participants of the European Conference on Digital Libraries (ECDL) 2006 and the on-line encyclopædia Wikipedia.

The test data to be used by the systems consisted of 30 person names from different sources: (i) 10 names were selected from Wikipedia; (ii) 10 names were selected from participants in the ACL 2006 conference; and finally, (iii) 10 further names were selected from the US Census. One hundred documents were retrieved using the person name as a query using the search engine Yahoo!.

Metrics used to measure the performance of automatic systems against the human output were borrowed from the clustering literature (Hotho et al., 2003) and they are defined as follows:

$$\text{Precision}(A, B) = \frac{|A \cap B|}{|A|}$$

$$\text{Purity}(C, L) = \sum_{i=1}^{n} \frac{|C_i|}{n} max_j \text{Precision}(C_i, L_j)$$

$$\text{Inverse\_Purity}(C, L) = \sum_{i=1}^{n} \frac{|L_i|}{n} max_j \text{Precision}(L_i, C_j)$$

$$\text{F-Score}_\alpha(C, L) = \frac{\text{Purity}(C, L) * \text{Inverse\_Purity}(C, L)}{\alpha \text{Purity}(C, L) + (1 - \alpha)\text{Inverse\_Purity}(C, L)}$$

where C is the set of clusters to be evaluated and L is the set of clusters produced by the human. Note that purity is a kind of precision metric which rewards a partition which has less noise. Inverse purity is a kind of recall metric. α was set to 0.5 in the SemEval 2007 evaluation. Two simple baseline systems were defined in order to measure if the techniques used by participants were able to improve over them. The all-in-one baseline produces one single cluster – all documents belonging to that cluster. The one-in-one baseline produces n cluster with one different document in each cluster.

3 Agglomerative Clustering Algorithm

Clustering is an important technique used in areas such as information retrieval, text mining, and data mining (Cutting et al., 1992). Clustering algorithms combine data points into groups such that: (i) data points in the same group are similar to each other; and (ii) data points in one group are "different" from data points in a different group or cluster. In information retrieval it is assumed that documents that are similar to each other are likely to be relevant for the same query, and therefore having the document collection organised in clusters can provide improved document access (van Rijsbergen, 1979). Different clustering techniques exist (Willett, 1988) the simplest one being the one-pass clustering algorithm (Rasmussen and Willett, 1987). We have implemented an agglomerative clustering algorithm which is relatively simple, has reasonable complexity, and gave us rather good results. Our algorithm operates in an exclusive way, meaning that a document belongs to one and only one cluster – while this is our working hypothesis, it might not be valid in some cases.

The input to the algorithm is a set of document representations implemented as vectors of terms and weights. Initially, there are as many clusters as input documents; as the algorithm proceeds clusters are merged until a certain termination condition is reached. The algorithm computes the similarity between vector representations in order to decide whether or not to merge two clusters.

The similarity metric we use is the cosine of the angle between two vectors. This metric gives value one for identical vectors and zero for vectors which are orthogonal (non related). Various options have been implemented in order to measure how close

two clusters are, but for the experiments reported here we have used the following approach: the similarity between two clusters (sim_C) is equivalent to the "document" similarity (sim_D) between the two more similar documents in the two clusters – this is known as single linkage in the clustering literature; the following formula is used:

$$\text{sim}_C(\text{C}_1,\text{C}_2) = \max_{d_i \in C_1; d_j \in C_2} \text{sim}_D(\text{d}_i,\text{d}_j)$$

Where C_k are clusters, d_l are document representations (e.g., vectors), and sim_D is the cosine metric given by the following formula:

$$\text{cosine}(d_1, d_2) = \frac{\sum_{i=1}^{n} w_{i,d_1} * w_{i,d_2}}{\sqrt{\sum_{i=1}^{n} (w_{i,d_1})^2} * \sqrt{\sum_{i=1}^{n} (w_{i,d_2})^2}}$$

where $w_{i,d}$ is the weight of term i in document d and n is the numbers of terms.

If this similarity is greater than a threshold – experimentally obtained – the two clusters are merged together. At each iteration the most similar pair of clusters is merged. If this similarity is less than a certain threshold the algorithm stops. Merging two clusters consist of a simple step of *set union*, so there is no re-computation involved – such as computing a cluster centroid.

We estimated the threshold for the clustering algorithm using the ECDL subset of the training data provided by SemEval. We applied the clustering algorithm where the threshold was set to zero. For each document set, purity, inverse purity, and F-score were computed at each iteration of the algorithm, recording the similarity value of each newly created cluster. The similarity values for the best clustering results (best F-score) were recorded, and the maximum and minimum values discarded. The rest of the values were averaged to obtain an estimate of the optimal threshold. The thresholds used for the experiments reported here are as follows: 0.10 for word vectors and 0.12 for named entity vectors (see Section 5 for vector representations).

4 Natural Language Processing Technology

We rely on available extraction and summarization technology in order to linguistically process the documents for creating document representations for clustering. Although the SemEval corpus contains information other than the retrieved pages themselves, we have made no attempt to analyse or use contextual information given with the input document.

Two tools are used: the GATE system (Cunningham et al., 2002) and a summarization toolkit (Saggion, 2002; Saggion and Gaizauskas, 2004) which is compatible with GATE. The input for analysis is a set of documents and a person name (first name and last name). The documents are analysed by the default GATE[1] ANNIE system which creates different types of named entity annotations. No adaptation of the system was carried out because we wanted to verify how far we could go using available tools. Summarization technology was used from single document summarization modules from our summarization toolkit.

The core of the toolkit is a set of summarization modules which compute numeric features for each sentence in the input document, the value of the feature indicates how relevant the information in the sentence is for the feature. The computed values, which are normalised yielding numbers in the interval [0..1] – are combined in a linear formula to obtain a score for each sentence which is used as the basis for sentence selection. Sentences are ranked based on their score and top ranked sentences selected to produce an extract. Many features implemented in this tool have been suggested in past research as valuable for the task of identifying sentences for creating summaries. In this work, summaries are created following two different approaches as described below.

The text and linguistic processors used in our system are: document tokenisation to identify different kinds of words; sentence splitting to segment the text into units used by the summariser; parts-of-speech tagging used for named entity recognition; named entity recognition using a gazetteer lookup module and regular expressions grammars; and named entity coreference module using a rule-based orthographic name matcher to identify name mentions considered equivalent (e.g., "John Smith" and "Mr. Smith"). Named entities of type *Person*, *Organization*, *Address*, *Date*, and *Location* are considered relevant

[1] `http://gate.ac.uk`

document terms and stored in a special named entity called *Mention* as an annotation. The performance of the named entity recogniser on Web data (business news from the Web) is around 0.90 F-score (Maynard et al., 2003).

Coreference chains are created and analysed and if they contain an entity matching the target person's surname, all elements of the chain are marked as a feature of the annotation.

We have tested two summarization conditions in this work: In one set of experiments a sentence belongs to a summary if it contains a mention which is coreferent with the target entity. In a second set of experiments a sentence belongs to a summary if it contains a "biographical pattern". We rely on a number of patterns that have been proposed in the past to identify *descriptive phrases* in text collections (Joho and Sanderson, 2000). The patterns used in the experiments described here are shown in Table 1. In the patterns, *dp* is a *descriptive phrase* that in (Joho and Sanderson, 2000) is taken as a noun phrase. These patterns are likely to capture information which is relevant to create person profiles, as used in DUC 2004 and in TREC QA – to answer definitional questions.

These patterns are implemented as regular expressions using the JAPE language (Cunningham et al., 2002). Our implementation of the patterns make use of coreference information so that *target* is *any* name in text which is coreferent with sought person. In order to implement the *dp* element in the patterns we use the information provided by a noun phrase chunker. The following is one of the JAPE rules for identifying key phrases as implemented in our system:

```
({TargetPerson}
({ Token.string == "is" } |
{Token.string == "was" })
{NounChunk}):annotate --> :annotate.KeyPhrase = {}
```

where *TargetPerson* is the sought entity, and *NounChunk* is a noun chunk. The rule states that when the pattern is found, a *KeyPhrase* should be created.

Some examples of these patterns in text are shown in Table 4. A profile-based summarization system which uses these patterns to create person profiles is reported in (Saggion and Gaizauskas, 2005).

Patterns
target (is | was |...) (a | an | the) dp
target, (who | whose | ...)
target, (a | the | one ...) dp
target, dp
target's
target and others

Table 1: Set of patterns for identifying profile information.

Dickson's invention, the Kinetoscope, was simple: a strip of several images was passed in front of an illuminated lens and behind a spinning wheel.

James Hamilton, 1st earl of Arran

James Davidson, MD, Sports Medicine Orthopedic Surgeon, Phoenix Arizona

As adjutant general, **Davidson was chief** of the State Police, qv which he organized quickly.

Table 2: Descriptive phrases in test documents for different target names.

4.1 Frequency Information

Using language resources creation modules from the summarization tool, two frequency tables are created for each document set (or person) on-the-fly: (i) an inverted document frequency table for *words* (no normalisation is applied); and (ii) an inverted frequency table for *Mentions* (the full entity string is used, no normalisation is applied).

Statistics (term frequencies (tf(Term)) and inverted document frequencies (idf(Term))) are computed over tokens and *Mentions* using tools from the summarization toolkit (see examples in Table 3).

word frequencies	Mention frequencies
of (92)	Jerry Hobbs (80)
Hobbs (92)	Hobbs (56)
Jerry (90)	Krystal Tobias (38)
to (89)	Texas (37)
in (87)	Jerry (36)
and (86)	Laura Hobbs (35)
the (85)	Monday (34)
a (85)	1990 (31)

Table 3: Examples of top frequent terms (words and named entities) and their frequencies in the Jerry Hobbs set.

Using these tables vector representations are created for each document (same as in (Bagga and

Baldwin, 1998)). We use the following formula to compute term weight (N is the number of documents in the input set):

$$\text{weight(Term)} = \text{tf(Term)} * log_2(\frac{N}{\text{idf(Term)}})$$

These vectors are also stored in the GATE documents. Two types of representations were considered for these experiments: (i) full document or summary (terms in the summary are considered for vector creation); and (ii) words are used as terms or *Mentions* are used as terms.

5 Cross-document Coreference Systems

In this section we present results of six different configurations of the clustering algorithm. The configurations are composed of two parts one which indicates where the terms are extracted from and the second part indicates what type of terms were used. The text conditions are as follows: *Full Document* (FD) condition means that the whole document was used for extracting terms for vector creation; *Person Summary* (PS) means that sentences containing the target person name were used to extract terms for vector creation; *Descriptive Phrase* (DP) means that sentences containing a descriptive patterns were used to extract terms for vector creation. The term conditions are: *Words* (W) words were used as terms and *Mentions* (M) named entities were used as terms. Local inverted term frequencies were used to weight the terms.

6 SemEval 2007 Web People Search Results

The best system in SemEval 2007 obtained an F-score of 0.78, the average F-score of all 16 participant systems is 0.60. Baseline *one-in-one* has an F-score of 0.61 and baseline *all-in-one* an F-score of 0.40. Results for our system configurations are presented in Table 4. Our best configuration (FD+W) obtains an F-score of 0.74 (or a fourth position in the SemEval ranking). All our configurations obtained F-scores greater than the average of 0.60 of all participant systems. They also perform better than the two baselines.

Our optimal configurations (FD+W and PS+W) both perform similarly with respect to F-score. While the full document condition favours "inverse purity", summary condition favours "purity". As one may expect, the use of descriptive phrases to create summaries has the effect of increasing purity to one extreme, these expressions are far too restrictive to capture all necessary information for disambiguation.

Configuration	Purity	Inv.Purity	F-Score
FD+W	0.68	0.85	0.74
FD+M	0.62	0.85	0.68
PS+W	0.84	0.70	0.74
PS+M	0.65	0.75	0.64
DP+W	0.90	0.62	0.71
DP+M	0.97	0.53	0.66

Table 4: Results for different clustering configurations. These results are those obtained on the whole set of 30 person names.

7 Semantic-based Experiments

While these results are rather encouraging, they were not optimal. In particular, we were surprised that semantic information performed worst than a simple word-based approach. We decided to investigate whether some types of semantic information might be more helpful than others in the clustering process. We therefore created one vector for each type of information: *Organization*, *Person*, *Location*, *Date*, *Address* in each document and reclustered all test data using one type at a time, without modifying any of the system parameters (e.g., without re-training). The results were very encouraging.

7.1 Results

Results of semantic-based clustering per information type are presented in Tables 5 and 6. Each row

Semantic Type	Purity	Inv.Purity	F-Score	+/-
Organization	0.90	0.72	0.78	+0.10
Person	0.81	0.72	0.75	+0.07
Address	0.82	0.64	0.69	+0.01
Date	0.58	0.85	0.67	-0.01
Location	0.55	0.85	0.64	-0.04

Table 5: Results for full document condition and different semantic information types. Improvements over FD+M are reported.

Semantic Type	Purity	Inv.Purity	F-Score	+/-
Person	0.85	0.64	0.70	+0.06
Organization	0.97	0.57	0.69	+0.05
Date	0.87	0.60	0.68	+0.04
Location	0.82	0.63	0.67	+0.03
Address	0.93	0.54	0.65	+0.01

Table 6: Results for summary condition and different semantic information types. Improvements over PS+M are reported.

in the tables reports results for clustering using one type of information alone. Table 5 reports results for semantic information with full text condition and it is therefore compared to our configuration FD+M which also uses full text condition together with semantic information. The last column in the table shows improvements over that configuration. Using *Organization* type of information in full text condition, not only outperforms the previous system by ten points, also exceeds by a fraction of a point the best system in SemEval 2007 (one point if we consider macro averaged F-score). Statistical tests (*t*-test) show that improvement over FD+M is statistically significant. Other semantic types of information also have improved performance, not all of them however. *Location* and *Date* in the full documents are probably too ambiguous to help disambiguating the target named entity.

Table 6 reports results for semantic information with summary text condition (only personal summaries were tried, experiments using descriptive phrases are underway) and it is therefore compared to our configuration PS+M which also uses summary condition together with semantic information. The last column in the table shows improvements over that configuration. Here all semantic types of information taken individually outperform a system which uses the combination of all types. This is probably because all types of information in a personal summary are somehow related to the target person.

7.2 Results per Person Set

Following (Popescu and Magnini, 2007), we present purity, inverse purity, and F-score results for all our configurations per category (ACL, US Census, Wikipedia) in the test set.

In Tables 7, 8, and 9, results are reported for full

Configuration	Set	Purity	I.Purity	F-Score
FD+Address	ACL	0.86	0.48	0.57
FD+Address	US C.	0.81	0.71	0.75
FD+Address	Wikip.	0.78	0.70	0.73
PS+Address	ACL	0.96	0.38	0.50
PS+Address	US C.	0.94	0.61	0.72
PS+Address	Wikip.	0.88	0.62	0.71
FD+Date	ACL	0.63	0.82	0.69
FD+Date	US C.	0.52	0.87	0.64
FD+Date	Wikip.	0.59	0.85	0.68
PS+Date	ACL	0.88	0.49	0.59
PS+Date	US C.	0.88	0.64	0.72
PS+Date	Wikip.	0.84	0.67	0.72
FD+Location	ACL	0.63	0.78	0.65
FD+Location	US C.	0.52	0.86	0.64
FD+Location	Wikip.	0.49	0.91	0.62
PS+Location	ACL	0.87	0.47	0.54
PS+Location	US C.	0.85	0.66	0.73
PS+Location	Wikip.	0.74	0.75	0.72

Table 7: Results for clustering configurations per person type set (ACL, US Census, and Wikipedia) - Part I.

Configuration	Set	Purity	I.Purity	F-Score
FD+Org.	ACL	0.92	0.57	0.69
FD+Org.	US C.	0.87	0.78	0.82
FD+Org.	Wikip.	0.88	0.79	0.83
PS+Org.	ACL	0.98	0.42	0.54
PS+Org.	US C.	0.95	0.63	0.74
PS+Org.	Wikip.	0.96	0.65	0.77
FD+Person	ACL	0.82	0.66	0.72
FD+Person	US C.	0.81	0.74	0.76
FD+Person	Wikip.	0.77	0.75	0.75
PS+Person	ACL	0.86	0.53	0.63
PS+Person	US C.	0.85	0.6721	0.73
PS+Person	Wikip.	0.82	0.70	0.73

Table 8: Results for clustering configurations per person type set (ACL, US Census, and Wikipedia) - Part II.

document condition(FD), summary condition (PS), word-based representation (W), mention representation (M) – i.e. all types of named entities, and five different mention types: Person, Location, Organization, Date, and Address.

While the Organization type of entity worked better overall, it is not optimal across different categories of people. Note for example that very good results are obtained for the Wikipedia and US Census sets, but rather poor results for the ACL set, where a technique which relies on using full documents and words for document representations works better. These results show that more work is

Configuration	Set	Purity	I.Purity	F-Score
FD+W	ACL	0.73	0.84	0.77
FD+W	US C.	0.54	0.91	0.67
FD+W	Wikip.	0.57	0.91	0.68
FD+M	ACL	0.73	0.76	0.70
FD+M	US C.	0.68	0.82	0.71
FD+M	Wikip.	0.60	0.86	0.68
PS+W	ACL	0.84	0.59	0.65
PS+W	US C.	0.80	0.74	0.75
PS+W	Wikip.	0.70	0.81	0.73
PS+M	ACL	0.75	0.62	0.60
PS+M	US C.	0.71	0.74	0.69
PS+M	Wikip.	0.58	0.83	0.66

Table 9: Results for clustering configurations per person type set (ACL, US Census, and Wikipedia) - Part III.

needed before reaching any conclusions on the best document representation for our algorithm in this task.

8 Related Work

The problem of cross-document coreference has been studied for a number of years now. Bagga and Baldwin (Bagga and Baldwin, 1998) used the vector space model together with summarization techniques to tackle the cross-document coreference problem. Their approach uses vector representations following a bag-of-words approach. Terms for vector representation are obtained from sentences where the target person appears. They have not presented an analysis of the impact of full document versus summary condition and their clustering algorithm is rather under-specified. Here we have presented a clearer picture of the influence of summary vs full document condition in the clustering process.

Mann and Yarowsky (Mann and Yarowsky, 2003) used semantic information extracted from documents referring to the target person in an hierarchical agglomerative clustering algorithm. Semantic information here refers to factual information about a person such as the date of birth, professional career or education. Information is extracted using patterns some of them manually developed and others induced from examples. We differ from this approach in that our semantic information is more general and is not particularly related - although it might be - to the target person.

Phan el al. (Phan et al., 2006) follow Mann and Yarowsky in their use of a kind of biographical information about a person. They use a machine learning algorithm to classify sentences according to particular information types in order to automatically construct a person profile. Instead of comparing biographical information in the person profile altogether as in (Mann and Yarowsky, 2003), they compare each type of information independently of each other, combining them only to make the final decision.

Finally, the best SemEval 2007 Web People Search system (Chen and Martin, 2007) used techniques similar to ours: named entity recognition using off-the-shelf systems. However in addition to semantic information and full document condition they also explore the use of contextual information such as the url where the document comes from. They show that this information is of little help. Our improved system obtained a slightly higher macro-averaged f-score over their system.

9 Conclusions and Future Work

We have presented experiments on cross-document coreference of person names in the context of the first SemEval 2007 Web People Search task. We have designed and implemented a solution which uses an in-house clustering algorithm and available extraction and summarization techniques to produce representations needed by the clustering algorithm. We have presented different approaches and compared them with SemEval evaluation's results. We have also shown that one system which uses one specific type of semantic information achieves state-of-the-art performance. However, more work is needed, in order to understand variation in performance from one data set to another.

Many avenues of improvement are expected. Where extraction technology is concerned, we have used an off-the-shelf system which is probably not the most appropriate for the type of data we are dealing with, and so adaptation is needed here. With respect to the clustering algorithm we plan to carry out further experiments to test the effect of different similarity metrics, different merging criteria including creation of cluster centroids, and cluster distances; with respect to the summarization techniques we intend to investigate how the extraction of sentences

containing pronouns referring to the target entity affects performance, our current version only exploits name coreference. Our future work will also explore how (and if) the use of contextual information available on the web can lead to better performance.

Acknowledgements

We are indebted to the three anonymous reviewers for their extensive suggestions that helped improve this work. This work was partially supported by the EU-funded MUSING project (IST-2004-027097).

References

J. Artiles, J. Gonzalo, and S. Sekine. 2007. The SemEval-2007 WePS Evaluation: Establishing a benchmark for 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 the 36th Annual Meeting of the Association for Computational Linguistics and the 17th International Conference on Computational Linguistics (COLING-ACL'98)*, pages 79–85.

Y. Chen and J.H. Martin. 2007. Cu-comsem: Exploring rich features for unsupervised web personal named disambiguation. In *Proceedings of SemEval 2007, Assocciation for Computational Linguistics*, pages 125–128.

H. Cunningham, D. Maynard, K. Bontcheva, and V. Tablan. 2002. GATE: A Framework and Graphical Development Environment for Robust NLP Tools and Applications. In *Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics (ACL'02)*.

Douglass R. Cutting, Jan O. Pedersen, David Karger, and John W. Tukey. 1992. Scatter/gather: A cluster-based approach to browsing large document collections. In *Proceedings of the Fifteenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 318–329.

A. Hotho, S. Staab, and G. Stumme. 2003. WordNet improves text document clustering. In *Proc. of the SIGIR 2003 Semantic Web Workshop*.

H. Joho and M. Sanderson. 2000. Retrieving Descriptive Phrases from Large Amounts of Free Text. In *Proceedings of Conference on Information and Knoweldge Management (CIKM)*, pages 180–186. ACM.

G. S. Mann and D. Yarowsky. 2003. Unsupervised personal name disambiguation. In W. Daelemans and M. Osborne, editors, *Proceedings of the 7^{th} Conference on Natural Language Learning (CoNLL-2003)*, pages 33–40. Edmonton, Canada, May.

D. Maynard, K. Bontcheva, and H. Cunningham. 2003. Towards a semantic extraction of named entities. In G. Angelova, K. Bontcheva, R. Mitkov, N. Nicolov, and N. Nikolov, editors, *Proceedings of Recent Advances in Natural Language Processing (RANLP'03)*, pages 255–261, Borovets, Bulgaria, Sep. `http://gate.ac.uk/sale/ranlp03/ranlp03.pdf`.

X.-H. Phan, L.-M. Nguyen, and S. Horiguchi. 2006. Personal name resolution crossover documents by a semantics-based approach. *IEICE Trans. Inf. & Syst.*, Feb 2006.

Octavian Popescu and Bernardo Magnini. 2007. Irst-bp: Web people search using name entities. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, pages 195–198, Prague, Czech Republic, June. Association for Computational Linguistics.

E. Rasmussen and P. Willett. 1987. Non-hierarchical document clustering using the icl distribution array processor. In *SIGIR '87: Proceedings of the 10th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 132–139, New York, NY, USA. ACM Press.

H. Saggion and R. Gaizauskas. 2004. Multi-document summarization by cluster/profile relevance and redundancy removal. In *Proceedings of the Document Understanding Conference 2004*. NIST.

H. Saggion and R. Gaizauskas. 2005. Experiments on statistical and pattern-based biographical summarization. In *Proceedings of EPIA 2005*, pages 611–621.

H. Saggion. 2002. Shallow-based Robust Summarization. In *Automatic Summarization: Solutions and Perspectives*, ATALA, December, 14.

H. Saggion. 2007. Shef: Semantic tagging and summarization techniques applied to cross-document coreference. In *Proceedings of SemEval 2007, Assocciation for Computational Linguistics*, pages 292–295.

C.J. van Rijsbergen. 1979. *Information Retrieval*. Butterworths, London.

P. Willett. 1988. Recent trends in hierarchic document clustering: A critical review. *Information Processing & Management*, 24(5):577–597.