SHEF: Semantic Tagging and Summarization Techniques Applied to 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 experiments for the cross-document coreference task in SemEval 2007. Our cross-document coreference system uses an in-house agglomerative clustering implementation to group documents referring to the same entity. Clustering uses vector representations created by summarization and semantic tagging analysis components. We present evaluation results for four system configurations demonstrating the potential of the applied techniques.

1 Introduction

Cross-document coreference resolution is the task of identifying if two mentions of the same (or similar) name in different sources refer to the same individual. Deciding if two documents refer to the same individual is a difficult problem because names are highly ambiguous. 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 and information extraction. Here, we concentrate on the following SemEval 2007 Web People Search Task (Artiles et al., 2007): a search engine user types in a person name as a query. Instead of ranking web pages, an ideal system should organize 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.

As past and recent research (Bagga and Baldwin, 1998; Phan et al., 2006), we have addressed the problem as a document clustering problem. For our first participation in SemEval 2007, we use two approaches: a lexical or bag-of-words approach and a semantic based approach. We have implemented our own clustering algorithms but rely on available extraction and summarization technology developed in our laboratory to produce document representations used as input for the clustering procedure.

2 Clustering Algorithm

We have implemented an agglomerative clustering algorithm. 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; the following formula is used:

Proceedings of the 4th International Workshop on Semantic Evaluations (SemEval-2007), pages 292–295,
Prague, June 2007. ©2007 Association for Computational Linguistics

$$\mathrm{sim}_C(\mathrm{C}_1,\mathrm{C}_2) =$$

$$\max_{d_i \in C_1; d_j \in C_2} \mathrm{sim}_D(\mathrm{d}_i,\mathrm{d}_j)$$

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

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.

3 Extraction and Summarization

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 (Cunningham et al., 2002) and single document summarization modules (Saggion and Gaizauskas, 2004b) from our summarization toolkit[2]. No attempt is made to analyse or use contextual information given with the input document. The processing elements include:

- Document tokenisation
- Sentence splitting
- Parts-of-speech tagging
- Named Entity Recognition using a gazetteer lookup module and regular expressions
- Named entity coreference using an orthographic name matcher

Named entities of type *person*, *organization*, *address*, *date*, and *location* are considered relevant document terms and stored in a special named entity called *Mention*.

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. Extractive summaries are created for each document, a sentence belongs to the summary if it contains a mention which is coreferent with the target entity.

Using language resources creation modules from the summarization tool, two frequency tables are created for each document set (or person): (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 and *tf*idf*) are computed over tokens and *Mentions* using the appropriate tables (these tools are part of the summarization toolkit) and vector representations created for each document (same as in (Bagga and Baldwin, 1998)). 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 or *Mentions*.

4 System Configurations

Four system configurations were prepared for SemEval:

- System I: vector representations were created for full documents. Words were used as terms and local inverted document frequencies used (word frequencies) for weighting.
- System II: vector representations were created for full documents. *Mentions* were used as terms and local inverted document frequencies used (Mentions frequencies) for weighting.
- System III: vector representations were created for person summaries. Words were used as terms and local inverted document frequencies used (word frequencies) for weighting.
- System IV: vector representations were created for person summaries. *Mentions* were used as terms and local inverted document frequencies used (Mentions frequencies) for weighting.

Because only one system configuration was allowed per participant team, we decided to select System II for official evaluation interested in evaluating the effect of semantic information in the clustering process.

5 Parameter Setting and Results

Evaluation of the task was carried out using standard clustering evaluation measures of "purity" and "inverse purity" (Hotho et al., 2003), and the harmonic

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

[2] http://www.dcs.shef.ac.uk/~saggion

Configuration	Purity	Inv.Purity	F-Score
System I	0.68	0.85	0.74
System II	0.62	0.85	0.68
System III	0.84	0.70	0.74
System IV	0.65	0.75	0.64

Table 1: Results for our configurations omitting one set. System II was the system we evaluated in SemEval 2007.

mean of purity and inverse purity: F-score. We estimated the threshold for the clustering algorithm using the ECDL subset of the training data provided by SemEval. We applied the clustering algorithm to each document set and computed purity, inverse purity, and F-score 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. Two different thresholds were obtained: 0.10 for word vectors and 0.12 for named entity vectors.

Results for the test set in SemEval are presented in Table 1 (One set – "Jerry Hobbs" – was ignored when computing these numbers: due to a failure during document analysis this set could not be clustered. The error was identified too close to the submission's date to allow us to re-process the cluster). Our official submission System II (SHEF in the official results) obtained an F-score of 0.66 positioning itself in 5th place (out of 16 systems). Our best configuration obtained 0.74 F-score, so a fourth place would be in theory possible.

Our system obtained an F-score greater than the average of 0.60 of all participant systems. Our optimal configurations (System I and System II) both perform similarly with respect to F-score. While System I favours "inverse purity", System III favours "purity". Results for every individual set are reported in the Appendix.

6 Conclusions and Future Work

We have presented a system used to participate in the SemEval 2007 Web People Search task. The system uses an in-house clustering algorithm and available extraction and summarization techniques to produce representations needed by the clustering algorithm. Although the configuration we submitted was suboptimal, we have obtained good results; in fact all our system configurations produce results well above the average of all participants. Our future work will explore how the use of contextual information available on the web can lead to better performance. We will explore if a similar approach to our method for creating profiles or answering definition questions (Saggion and Gaizauskas, 2004a) which uses co-occurence information to identify pieces of information related to a given entity can be applied here.

Acknowledgements

This work was partially supported by the EU-funded MUSING project (IST-2004-027097) and the EU-funded LIRICS project (eContent project 22236).

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.

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)*.

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

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.

H. Saggion and R. Gaizauskas. 2004a. Mining on-line sources for definition knowledge. In *Proceedings of the 17th FLAIRS 2004*, Miami Bearch, Florida, USA, May 17-19. AAAI.

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

Appendix I: Detailed Results

The following tables present purity, inverse purity, and F-score results for all sets and systems. These results were computed after re-processing the "Jerry Hobbs" missing set.

	System I			System II		
Person	**Pur.**	**I-Pur.**	**F**	**Pur.**	**I-Pur.**	**F**
Alvin Cooper	0.72	0.87	0.79	0.86	0.70	0.77
Arthur Morgan	0.90	0.83	0.86	0.75	0.92	0.83
Chris Brockett	0.87	0.85	0.86	0.94	0.67	0.78
Dekang Lin	1.00	0.63	0.77	1.00	0.66	0.79
Frank Keller	0.68	0.81	0.74	0.65	0.66	0.66
George Foster	0.61	0.83	0.71	0.45	0.88	0.60
Harry Hughes	0.82	0.80	0.81	0.71	0.93	0.80
James Curran	0.76	0.74	0.75	0.53	0.84	0.65
James Davidson	0.74	0.91	0.82	0.59	0.90	0.71
James Hamilton	0.52	0.90	0.66	0.25	0.97	0.39
James Morehead	0.38	0.91	0.54	0.39	0.92	0.55
Jerry Hobbs	0.67	0.86	0.75	0.61	0.85	0.71
John Nelson	0.64	0.93	0.76	0.56	0.90	0.69
Jonathan Brooks	0.70	0.89	0.78	0.54	0.89	0.67
Jude Brown	0.75	0.80	0.78	0.74	0.77	0.75
Karen Peterson	0.60	0.92	0.72	0.19	1.00	0.32
Leon Barrett	0.75	0.84	0.80	0.43	0.96	0.59
Marcy Jackson	0.60	0.91	0.72	0.87	0.85	0.86
Mark Johnson	0.57	0.86	0.68	0.33	0.94	0.49
Martha Edwards	0.49	0.96	0.65	0.43	0.91	0.58
Neil Clark	0.74	0.83	0.78	0.60	0.76	0.67
Patrick Killen	0.83	0.77	0.80	0.82	0.77	0.79
Robert Moore	0.64	0.78	0.71	0.44	0.91	0.60
Sharon Goldwater	1.00	0.80	0.89	1.00	0.80	0.89
Stephan Johnson	0.84	0.87	0.85	0.97	0.69	0.81
Stephen Clark	0.63	0.87	0.73	0.57	0.83	0.67
Thomas Fraser	0.51	0.94	0.66	0.44	0.94	0.60
Thomas Kirk	0.66	0.94	0.78	0.87	0.92	0.90
Violet Howard	0.34	0.96	0.51	0.71	0.90	0.80
William Dickson	0.55	0.94	0.70	0.38	0.95	0.54
AVERAGES	0.68	0.86	0.74	0.62	0.85	0.68

	System III			System VI		
Person	**Pur.**	**I-Pur.**	**F**	**Pur.**	**I-Pur.**	**F**
Alvin Cooper	0.98	0.58	0.73	0.93	0.52	0.67
Arthur Morgan	0.98	0.64	0.78	0.71	0.79	0.75
Chris Brockett	1.00	0.32	0.49	0.95	0.31	0.47
Dekang Lin	1.00	0.40	0.58	1.00	0.34	0.51
Frank Keller	0.85	0.65	0.74	0.50	0.71	0.59
George Foster	0.80	0.80	0.80	0.48	0.86	0.61
Harry Hughes	0.91	0.65	0.76	0.76	0.77	0.77
James Curran	0.92	0.69	0.79	0.64	0.77	0.70
James Davidson	0.82	0.85	0.83	0.48	0.93	0.63
James Hamilton	0.65	0.87	0.74	0.26	0.96	0.41
James Morehead	0.66	0.73	0.70	0.57	0.70	0.63
Jerry Hobbs	0.67	0.82	0.74	0.63	0.86	0.73
John Nelson	0.80	0.78	0.79	0.52	0.92	0.66
Jonathan Brooks	0.84	0.85	0.85	0.55	0.86	0.67
Jude Brown	0.75	0.72	0.74	0.80	0.69	0.74
Karen Peterson	0.80	0.86	0.83	0.26	0.94	0.41
Leon Barrett	0.91	0.52	0.66	0.79	0.62	0.69
Marcy Jackson	0.95	0.58	0.72	0.98	0.57	0.72
Mark Johnson	0.76	0.84	0.80	0.44	0.90	0.60
Martha Edwards	0.78	0.85	0.81	0.57	0.87	0.69
Neil Clark	0.85	0.53	0.65	0.60	0.75	0.67
Patrick Killen	0.99	0.57	0.73	0.90	0.61	0.73
Robert Moore	0.74	0.67	0.71	0.49	0.85	0.62
Sharon Goldwater	1.00	0.15	0.26	1.00	0.23	0.37
Stephan Johnson	0.94	0.71	0.81	0.95	0.71	0.81
Stephen Clark	0.87	0.80	0.83	0.55	0.82	0.66
Thomas Fraser	0.62	0.89	0.73	0.47	0.92	0.62
Thomas Kirk	0.81	0.87	0.84	0.84	0.86	0.85
Violet Howard	0.89	0.78	0.83	0.87	0.75	0.81
William Dickson	0.68	0.88	0.77	0.52	0.88	0.66
AVERAGES	0.84	0.70	0.73	0.67	0.74	0.65