The News People Search Task at EVALITA 2011: Evaluating Cross-document Coreference Resolution of Named Person Entities in Italian News

Luisa Bentivogli[1], Alessandro Marchetti[2], Emanuele Pianta[1]

[1]Fondazione Bruno Kessler
Trento, Italy
[2]CELCT
Trento, Italy
{bentivo, pianta}@fbk.eu, amarchetti@celct.it

Abstract. This paper describes the News People Search (NePS) Task organized as part of EVALITA 2011. The NePS Task aims at evaluating cross-document coreference resolution of person entities in Italian news and consists of clustering a set of Italian newspaper articles that mention a person name according to the different people sharing the name. The motivation behind the task, the dataset used for the evaluation and the results obtained are described and discussed.

1 Introduction: Motivations for the NePS Task

The News People Search Task (NePS) aims at evaluating cross-document coreference resolution of named person entities in Italian news. Cross-document coreference resolution consists in recognizing when different documents are referring to the same entity and represents a natural central component for a broad range of advanced NLP applications addressing multi-document processing, such as multi-document summarization, question answering, information extraction, entity detection and tracking, knowledge base population from texts.

In recent years, various initiatives such as the ACE 2008 [1] and WePS [2,3,4] evaluation campaigns made large annotated resources available and introduced quantitative evaluation of the cross-document coreference resolution task, allowing remarkable advances within the field. However, while such efforts are stimulating research for the English language, little has been done for other languages.

The NePS Task organized at EVALITA 2011 constitutes our contribution to the field of cross-document coreference resolution for the Italian language, by offering both a large annotated dataset and a common evaluation framework for cross-document coreference resolution systems working on Italian.

The paper is structured as follows. Section 2 describes the task, Section 3 presents a description of the dataset used for the evaluation, Section 4 introduces the evaluation measures used, Section 5 reports the results of the participating system, and finally Section 6 draws some conclusions about the evaluation exercise.

2   Definition of the Task

Cross-document coreference of a person entity occurs when the same person is mentioned in more than one text source. It can be defined as a clustering problem, which in principle requires the clustering of name occurrences in a corpus according to the persons they refer to. In the NePS Task, we consider clusters of *documents* containing the name occurrences. Cross-document coreference involves two problematic aspects, namely (i) to resolve ambiguities between people having the same name (i.e. when identical mentions refer to distinct persons) and, conversely, (ii) to recognize when different names refer to the same person.

The cross-document coreference resolution task has close links with Word Sense Disambiguation, which consists of deciding the sense of a word in a given context. In both tasks, the problem addressed is the resolution of the ambiguity in a natural language expression. More precisely, the NePS task can be viewed as a case of Word Sense Discrimination, as the number of "senses" (i.e. actual people carrying the same name) is unknown a priori.

The NePS task consists of clustering a set of Italian newspaper articles that mention a person name according to the different people sharing the name (i.e. one cluster of documents for each different person). More specifically, for each person name, systems receive in input a set of newspaper articles and the expected output is a clustering of the documents, where each cluster is supposed to contain all and only those documents that refer to the same individual.

The NePS task is limited to documents in which the entities are mentioned by name and takes into account name variability. Different kinds of name variants are considered, such as complete names (Paolo Rossi, Rossi Paolo), abbreviations (P. Rossi, Paolo R.), first names only (Paolo), last names only (Rossi), nicknames (Pablito), and misspellings (Paalo Rossi).

The scenario in which the task can be situated is that of an advanced search engine allowing intelligent access to newspaper information. In such scenario, an hypothetical user types a person name as a query and is presented with a set of clusters, where each cluster represents a specific entity and is assumed to contain all and only the newspaper articles referring to such entity.

The NePS task is structured along the same lines as the Web People Search evaluation exercise (WePS), which in 2010 was at its third edition. The main differences with respect to the WePS clustering task are that the NePS task (i) addresses Italian language instead of English, (ii) takes into account name variability, and (iii) uses a corpus of newspaper articles instead of web pages.

3   Dataset Description

The dataset used for the NePS task is the Cross-document Italian People Coreference corpus (CRIPCO). The CRIPCO corpus is composed of 43,328 documents representing a subset of the news stories published by the local newspaper "L'Adige" from 1999 to 2006.

The dataset was created selecting a representative number of person names (Group Names) as seed for the annotation of the corpus. Among all the possible name variants, a Group Name is always a complete name, i.e. a pair First Name-Last Name (e.g. Paolo Rossi, Isabella Bossi Fedrigotti, Diego Armando Maradona). For each Group Name, a number of documents containing at least one mention of the group name (or of one of its possible variants) were selected and clustered according to the actual person they refer to (i.e. one cluster of documents for each different person).

A detailed description of the principles upon which the corpus was created can be found in [5], whereas Table 1 presents information about its composition, also considering Development Set and Test Set separately.

Table 1. Corpus composition

	# Group Names	# Entities	# Documents
Development Set	105	342	22,574
Test Set	103	355	20,754
All Corpus	208	697	43,328

As for the average Group Name ambiguity in the dataset, it amounts to 3.35, meaning that on average 3.35 different persons (entities) share the same Group Name.

Given that the difficulty of the automatic coreference task varies on the basis of the ambiguity of the Group Name (i.e. the more ambiguous the Group Name, the more difficult is to disambiguate it), we subdivided the Group Names into three different ambiguity ranges, namely:

- no ambiguity: only one person carries the Group Name.
- medium ambiguity: from two to three persons share the same Group Name.
- high ambiguity: more than three persons share the same Group Name.

Table 2 presents the breakdown of the ambiguity of the Group Names in the dataset according to the three ambiguity ranges identified, together with a further subdivision into Development Set and Test Set.

Table 2. Distribution of Group Names ambiguity according to ambiguity ranges

	No Ambiguity			Medium Ambiguity			High Ambiguity		
	Dev	Test	*All*	Dev	Test	*All*	Dev	Test	*All*
# Group Names	51	48	*99*	23	24	*47*	31	31	*62*
# Entities	51	48	*99*	55	55	*110*	236	252	*488*
Average Ambiguity	1	1	*1*	2.391	2.292	***2.340***	7.613	8.129	***7.871***

4   Evaluation Measures

System results were compared to the human-annotated gold standard and the metrics used to evaluate system performances were Extended B-Cubed Precision and Recall [6], combined with F1 measure. The extended version of B-Cubed was introduced in the WePs-2 task to specifically address the evaluation of overlapping clustering: in case of non-overlapping clustering, extended B-Cubed results are identical to those obtained using standard B-Cubed. The evaluation was carried out using the official scorer distributed by the WePS orgainizers for the WePS-2 task[1], and runs were officially ranked according to their B-Cubed F1 score.

5   Participation Results

Five teams registered to the NePS Task and one team participated in it submitting one run. The system results have been compared to the so-called "ALL_IN_ONE" baseline, which considers all the documents of a given Group Name pertaining to the same person, thus giving the highest possible recall score.

The results of the evaluation are shown in Table 3. They are presented as overall scores on the whole Test Set, as well as grouped according to the ambiguity range of the Group Names.

---

[1]   http://nlp.uned.es/weps/weps-1/weps1-data

Table 3. Evaluation Results

	ALL			No ambiguity			Medium ambiguity			High ambiguity		
	BEP	BER	F1	BEP	BER	F1	BEP	BER	F1	BEP	BER	F1
FBK_0	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

Considering the overall F1 results, FBK_0 is 0.02 points above the ALL-IN-ONE baseline. It must be noticed that in this dataset the ALL-IN-ONE baseline results to be very high due to a number of factors. First, there is a high number of unambiguous names. Moreover, in the "highly ambiguous" category, the distribution of documents among the entities is skewed, as most of the documents refer to one single (usually famous) person carrying the ambiguous name.

As regards the breakdown of results according to the Group Name ambiguity, for Group Names with no ambiguity and medium ambiguity the results of the system lie very close to the baseline. The main differences with respect to the baseline can be noticed for the highly ambiguous names, where FBK F1 score is 0.10 above the baseline. It has to be pointed out that the system performs better than the baseline in the most interesting and difficult case.

6   Conclusions

In this paper we presented an evaluation task devoted to cross-document coreference resolution. The participating system showed good performances in the Test Set, performing similarly to the baseline considering the overall results, and 0.10 points above it when considering highly ambiguous Group Names.

With the NePS task and the CRIPCO corpus we tried to fill the gap of availability of annotated resources for cross-document coreference resolution for Italian language. As the availability of annotated data is crucial for advancing the state of the art in the field, we hope that the resource and the evaluation exercise proposed will help researchers to improve their systems and will encourage them to participate to future evaluation campaigns.

References

1. ACE 2008, http://projects.ldc.upenn.edu/ace/docs/ACE08_XDOC_1.6.pdf
2. Artiles, J., Gonzalo, J., and Sekine, S.: "The SemEval-2007 WePS Evaluation: Establishing a benchmark for the Web People Search Task". In: Proceedings of SemEval-2007, Prague, Czech Republic (2007)
3. Artiles, J., Gonzalo, J., and Sekine, S.: "WePS 2 Evaluation Campaign: Overview of the Web People Search Clustering Task". In: Proceedings of WePS 2 Workshop, Madrid, Spain (2009)

4. Artiles J., Borthwick A., Gonzalo J., Sekine S., and Amigó E.: "WePS-3 Evaluation Campaign: Overview of the Web People Search Clustering and Attribute Extraction Tasks". In: CLEF 2010 LABs and Workshops Notebook Papers, Padua, Italy (2010)
5. Bentivogli L., Girardi C., and Pianta E.: "Creating a Gold Standard for Person Cross-Document Coreference Resolution in Italian News". In: Proceedings of the LREC 2008 Workshop on Resources and Evaluation for Identity Matching, Entity Resolution and Entity Management. Marrakech, Morocco (2008)
6. Amigó, E., Gonzalo, J., Artiles, J., and Verdejo F.: A comparison of extrinsic clustering evaluation metrics based on formal constraints. Information Retrieval (2008)