

Creating a Gold Standard for Person Cross-Document Coreference Resolution in Italian News

Luisa Bentivogli, Christian Girardi, Emanuele Pianta

Fondazione Bruno Kessler (FBK-irst)

Via Sommarive 18 - POVO, 38100 Trento - Italy

E-mail: bentivo@fbk.eu, cgirardi@fbk.eu, pianta@fbk.eu

Abstract

This paper presents work aimed at the realization of a gold standard for cross-document coreference resolution of person entities in a corpus of Italian news. The gold standard has been created selecting a number of person names occurring in Adige-500K, a corpus composed of all the news stories published by the local newspaper "L'Adige" from 1999 to 2006. The corpus consists of 535,000 news stories, for a total of around 200 million tokens. To sample the person names in the corpus, we identified two dimensions, corresponding to two phenomena we intended to study, namely (i) the fame of the person entities and (ii) the ambiguity of person names. The first version of the gold standard is composed of 209 person names corresponding to 709 entities, for a total of 43,704 annotated documents.

1. Introduction

Recent years have seen an increase in the demand of content-annotated resources for new tasks in Natural Language Processing, such as content extraction and coreference resolution. Content extraction refers to the extraction of entities (e.g. persons, locations, organizations) and of relations between entities (e.g. affiliation of a person to an organization), while coreference analysis is the process of determining whether or not different text portions refer to the same entity. Entity coreference can be found both within the same document (intra-document), and across different documents in a corpus (cross-document).

Various initiatives such as MUC, ACE, SemEval made large annotated resources available and introduced quantitative evaluation, allowing remarkable advances within the fields of intra- and cross-document coreference. However, while such efforts are stimulating research for the English language, little has been done for other languages, where these kinds of resources are still lacking.

This paper constitutes our contribution to the field of cross-document coreference resolution, presenting a work aimed at the creation of a gold standard for cross-document coreference resolution of named person entities in Italian news.

This work has been carried out in the context of the OntoText (From Text to Knowledge for the Semantic Web) project, which has been funded by the Autonomous Province of Trento under the FUP-2004 research program. Based on the philosophy of the Semantic Web, OntoText exploits text processing and automatic reasoning technologies to extract knowledge from texts and organize it conceptually in an ontology. The new OntoText technologies have been applied and tested on the Italian corpus *Adige-500K*, which contains the news stories published by the local newspaper "L'Adige" from 1999 to 2006. The corpus consists of 535,000 news stories, for a total of around 200 million tokens. One of the main outcomes of the project is represented by the OntoText Portal, which provides an integrated access to the information automatically extracted from *Adige-500K*. Differently from common text-based search engines, the OntoText Portal directly accesses the

concepts and entities of the ontology and presents the user with structured information instead of mere portions of texts. As specifically regards entities of type PERSON, when an OntoText Portal user types a person name as a query, he/she 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.

One of the first uses of the gold standard is the evaluation of the coreference algorithm in charge of clustering the newspaper articles of the *Adige-500K* corpus according to the query of the OntoText Portal user.

The paper is structured as follows. Section 2 reports on other existing resources for cross-document coreference evaluation. Section 3 describes in detail the creation of the gold standard: its design, the annotation process, and all the data about its composition up to now. Section 4 presents the web interface specifically developed for the cross-document coreference annotation task. Finally, Section 5 draws some conclusions and explains future work.

2. Related Work

While intra-document coreference is a long dating and established area of research (e.g. anaphora resolution), the work on cross-document coreference resolution¹ began more recently. Bagga and Baldwin (1998) created the first reference data set for benchmarking cross-document coreference results, the *John Smith Corpus*, composed of 197 articles from the New York Times containing the name "John Smith". The John Smith corpus allows for evaluating only a subset of the cross-document entity coreference functionality as documents containing name variations of "John Smith" are not included.

The field has seen a rapidly growing interest (Mann and Yarowsky 2003, Gooi and Allan 2004, Blume 2005, Bollegala et al. 2006), however the algorithms for coreference resolution were generally evaluated on very

¹In the literature, this task is also referred to, with slight different meanings, as cross-document/interdocument/global coreference resolution, entity disambiguation, identity resolution.

few names in small corpora, or on artificial corpora, or through a posteriori control.

The first large size gold standard, which up to now represents the state of the art, has been created for the first cross-document coreference evaluation campaign, namely the SemEval-2007 Web People Search task (Artiles et al., 2007). The Web People Search corpus includes documents about 79 complete person names (first name and last name) corresponding to 1,882 entities mentioned in about 7,900 web pages (the 100 top results for a person name query to the Yahoo! search engine). The Web People Search corpus does not include documents with name variants and thus does not allow for name variation evaluation.

The resource that is most similar to our work is the forthcoming evaluation corpus of the ACE 2008 Global Entity Detection and Disambiguation task (ACE 2008), whose guidelines represent the standard to which we adhere. The ACE 2008 task consists in cross-document entity disambiguation, limited to documents in which entities are mentioned by name, be it the exact name or a name variant (e.g. long and short form of the name, variant spellings, misspellings, transliterations, aliases, and nicknames). According to the task, in the gold standard only coreference between named entities will be annotated.

The ACE 2008 corpus contains English and Arabic texts and will be composed of 10,000 documents per language. Only a subset of the whole corpus will be annotated for cross-document coreference.

As said above, all the resources available to the community up to now are for English. The only Italian resource annotated with cross-document coreference is the Italian Content Annotation Bank (I-CAB). I-CAB (Magnini et al., 2006) consists of 525 news documents taken from the local newspaper “L’Adige” for a total of around 182,500 words. The selected news stories belong to four different days (September, 7th and 8th 2004 and October, 7th and 8th 2004). The annotation of I-CAB has been carried out manually within the OntoText project, following the ACE annotation guidelines for the Entity Detection task, slightly modified to cope with the different morpho-syntactic characteristics of Italian. I-CAB is annotated with temporal expressions and with four types of entities, namely PERSON, ORGANIZATION, GEO-POLITICAL and LOCATION. Manual intra-document coreference has been carried out for all the annotated entities, with Callisto. Moreover, for PERSON and LOCATION entities also cross-document coreference has been carried out.

However, I-CAB is not suitable enough for evaluating cross-document coreference resolution as the newspaper articles have been chosen within a short time-span where very few different mentions of the same entity are found.

3. Creating the Gold Standard

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 this work, as in SemEval, 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 gold standard described in this paper addresses the annotation of cross-document coreference of named person entities in an Italian newspaper corpus.

The documents of the gold standard are selected from the Italian Adige-500K corpus. Given the number of documents in the corpus (more than 500,000) and the time-span covered (7 years), we think that Adige-500K is suitable for evaluation (and possibly training) of cross-document coreference, allowing for a great variety of name mentions and for entities to occur in a lot of documents.

Following the ACE 2008 guidelines, the annotation is limited to documents in which the entities are mentioned by name². Different kinds of name variants are considered, such as complete names (Paolo Rossi), abbreviations (P. Rossi, Paolo R.), first names only (Paolo), last names only (Rossi), nicknames (Pablito), and also misspellings (Paolo Rossi) and journalist errors in reporting the correct name of the entity (Carlo Rossi instead of Paolo Rossi).

A representative number of names occurring in the Adige-500K corpus have been selected as seeds for the creation of the gold standard (*Seed Names*). Among all the possible name variants, we decided that a Seed Name is always a complete name, i.e. a pair First Name-Last Name (e.g. Paolo Rossi, Isabella Bossi Fedrigotti, Diego Armando Maradona)³.

In order to select the Seed Names to be annotated, two main criteria have been adopted, corresponding to two phenomena that we intended to study. These criteria are discussed in the next section.

3.1. Gold Standard Design Criteria

The first issue to be addressed when creating the gold standard is how to sample the Seed Names.

Two dimensions have been selected, namely (i) the fame of the entities and (ii) the ambiguity of the Seed Names. The first dimension, which refers to the entity level, is strictly related to the context of application of the OntoText project, within which the gold standard has been created; the ambiguity dimension, which refers to the Seed Name level, is inherent in the cross-document coreference resolution task.

3.1.1. Entity Fame

Within the OntoText project we stressed the importance of the application context of the technologies developed, i.e. the OntoText Portal. We want to choose Seed Names

² In terms of the ACE categories, the entities considered are of type “PER”, subtype “Individual” and class “SPC” (i.e. a particular, specific and unique entity) while the mention type is “NAM” (i.e. a proper name reference to the entity).

³In a complete OntoText application scenario, the Seed Names should represent all the possible user’s query, i.e. all the name variant types. In the current version of the gold standard, we have introduced the restriction that the Seed Name is always a complete name but we are planning to add new Seed Names corresponding to other name variants.

which are representative of the OntoText Portal user queries.

We do not have data about actual user queries yet. However, we foresee that fame will be an important criterion to classify user queries. A great part of the user queries will be related to famous persons (which thus need to be adequately sampled in the corpus); however the user is likely to be asking information also about persons he/she knows, but who are not famous. For this reason we decided to include in the gold standard people belonging to five fame categories:

- Not famous
- Quite famous at the regional level
- Quite famous at the national level
- Very famous at the regional level
- Very famous at the national level

The distinction between the regional and the national level comes from the fact that the newspaper “L’Adige” contains both a national and a local section.

3.1.2. Name Ambiguity

The difficulty of the automatic coreference task varies on the basis of the ambiguity of the Seed Name: the more ambiguous the Seed Name, the more difficult is to disambiguate it. We want to study three different ambiguity scenarios:

- Low ambiguity
- Medium ambiguity
- High ambiguity

Summing up, we wanted the corpus to be structured along the two orthogonal variables of entity fame and Seed Name ambiguity. The original design of the gold standard is shown in Table 1, which partitions the expected set of entities of the gold standard in 15 cells; each cell is illustrated by the name of a sample entity.

	Not famous	Quite famous regional	Quite famous national	Very famous regional	Very famous national
Very ambiguous	Paolo Rossi	Elena Marino	Paolo Rossi		Paolo Rossi
Ambiguous	Franco Marini	Vittorio Colombo	Giovanna Marini		Franco Marini
Not ambiguous	Bruno Kessler	Dante Clauser	Marta Russo	Bruno Kessler	Umberto Eco

Table 1. Original design of the gold standard

In the original design, each cell in the grid was to be populated with entities, randomly selected from the Adige-500K corpus. However, to be able to use the standard evaluation techniques which are based on groups of entities carrying the same name (or variants of it), we decided that when we select an entity carrying a certain Seed Name for one cell, we also consider in the gold standard all other entities carrying the same Seed Name. Each time a given entity is introduced in the gold standard, also the other entities carrying the same name

are introduced. This makes a full balancing of the gold standard difficult to achieve.

Moreover, as hinted by the two empty cells in Table 1, some cells are intrinsically scarcely populated, namely those containing entities very famous at the regional level and carrying ambiguous names. This is explained by the fact that, in general, there are much more unambiguous names than ambiguous ones. Even rarer are the ambiguous names which occur in the corpus and refer to famous persons. All these constraints make the task of populating the “famous” class difficult, especially in a regional context, which is more restricted than the national one.

Thus we gave up the idea of a full balancing of the two variables (which implies selecting the same number of entities for each cell), and we decided to have all the classes of ambiguity and all the classes of fame populated with a minimum number of entities, which has been fixed at 30.

3.2. Selecting Seed Names

Starting from the list of all the 592,000 Named Entities of type PERSON automatically recognized in Adige-500K, we created a list of gold standard candidates by selecting those Named Entities (i) composed of at least two words, (ii) occurring at least in five different newspaper articles, and (iii) occurring in no more than 1,000 newspaper articles.

The first constraint is necessary in order to obtain a complete Seed Name, which is composed of first name and last name. The second constraint has been adopted to obtain entities interesting from the point of view of the cross-document coreference. The third was adopted for the practical reason that manually annotating more than 1,000 documents for one single Seed Name is too time expensive and error-prone.

From the resulting list of 79,000 gold standard candidates, we randomly picked up Seed Names until we found those satisfying our sampling criteria. At the end of the process, we had selected 209 Seed Names. The rules followed in the selection are described in the next sections.

3.2.1. Entity Fame

As regards the entity fame dimension, the first problem to face was how to evaluate the fame of a given entity.

To this purpose, we selected a pool of people of different ages and we asked them whether they had heard about some proposed entities, identified by a complete name and a short description. Then, we used the answers to classify the entities in the five categories described above.

Table 2 shows the distribution of the 209 selected Seed Names over the five fame categories. It is important to notice that at this preliminary stage of the gold standard creation we could work only at the Seed Name level. This is due to the fact that the knowledge about the actual different entities corresponding to a given Seed Name is not available a priori but only at the end of the coreference resolution process. Thus, the “famous” cells of Table 2 contain Seed Names for which we knew that there was at least one famous entity, whereas the “not famous” cell contains Seed Names not referring to any famous entity. Section 4.1. reports data about the actual

entity fame, obtained after manual Seed Name disambiguation.

Not famous	Quite famous Regional	Quite famous National	Very famous Regional	Very famous National
59	42	38	23	47

Table 2. Population of the fame categories

As it can be seen in Table 2, the category of very famous people at the regional level is populated with only 23 Seed Names instead of 30. This is due to the fact that, given the nature of the newspaper, almost all the people which are very famous at the regional level carry a name which belongs to the group of the top frequency names having more than 1,000 occurrences in the corpus, which were previously excluded from the gold standard.

3.2.2. Name Ambiguity

In order to evaluate the ambiguity of a Seed Name, we resorted to an external source (see Artiles et al., 2007). The source used is PagineBianche, the Italian telephone directory. We exploited the information related to the number of subscribers having the same name to create three ambiguity classes according to the thresholds reported in Table 3. Then the three classes were as much as possible equally populated. Table 3 also reports how the 209 Seed Names selected were grouped with respect to PagineBianche. The class of unambiguous Seed Names is much more populated than the other two classes. This is due to the fact that almost all the Seed Names selected in order to populate the classes of famous entities (first sampling criterion) belong to the class of unambiguous Seed Names.

Ambiguity	Number of subscribers	Selected Seed Names
Not ambiguous	0-99	121
Ambiguous	100-199	42
Very ambiguous	200+	46

Table 3. Population of the ambiguity classes

PagineBianche is the only large scale representation of the Italian population that we could find. Unfortunately, such representation is not totally accurate. The subscribers of PagineBianche are usually adult males permanently living with their family. Young people who only own mobile phones are not present in PagineBianche and the same is for the majority of women because the PagineBianche subscriber is usually their male partner. We decided to normalize the occurrences of women names multiplying by five the number of female names found in PagineBianche.

3.2.3. Number of Documents

Another dimension of the corpus that we tried to keep under control (by a posteriori check) is the number of documents where a Seed Name occurs, as this can have an influence on the difficulty of the coreference resolution task. The frequency range fixed a priori goes

from 5 to 1,000 occurrences of a given Seed Name in different newspaper articles. The number of documents containing the selected Seed Names cover most of this frequency range, with a minimum of 5 documents per Seed Name and a maximum of 893.

These ranges represent an approximation of the final number of documents associated to each Seed Name, as in this phase of the project the data about name variation of the Seed Names are not available yet.

The expected minimum size of the gold standard amounts to 32,582 Adige-500K documents, that is the number of documents containing a Seed Name mention.

Information about the correlation between the criteria according to which the gold standard has been modelled and the actual corpus data, which cannot be known a priori but only once the annotation has been carried out, will be given in Section 4.

3.3. The Annotation Process

To carry out the gold standard annotation, five annotators were selected and trained.

Given a certain Seed Name, the annotators have to disambiguate all the entities carrying that name and, for each entity, to find all the newspaper articles in which such entity is mentioned, both with its Seed Name and with all its possible name variants.

In this phase of the project, the annotators annotate the documents in which an entity is mentioned (in all its possible variants), but they do not annotate the single mentions of the entity within the documents.

In order to find all the possible name variants, the annotators can rely on a “lexicographer toolbox” (Giuliano, 2002) containing both concordances and collocations for the Adige-500K corpus. The toolbox turned out to be especially useful to find short forms of the names and misspellings.

The name variants found are used (together with any contextual word sequence identifying the entity) to create queries to the corpus, queries aimed at finding all the documents referring to the entity under consideration.

At the end of the annotation process, for each entity the result is (i) the identification of the documents referring to the entity and (ii) the creation of a list of its name variants.

According to the annotation guidelines, annotators are requested to take into consideration only entity mentions of type “proper name”. In some cases the documents should not be annotated because:

- The entity is not mentioned with a proper name. This is the case of entity descriptions (e.g. “Il Sindaco di Trento”/ “The Mayor of Trento” for the entity “Alberto Pacher”)
- The Seed Name refers to a non-person entity (e.g. organizations, streets, buildings named after a person, person names within titles of books, songs, etc.)
- The proper name refers to the author of the newspaper article.

In the case of non-informative documents, they are assigned to a “catch all” cluster. This happens for those documents containing only lists of names without any kind of further information.

As for the type of document annotation, the annotation can be marked as “not sure” in those cases where the

annotator is not sure if a document is referring to a specific entity or not.

In those cases where the same document refers to more than one entity carrying the same Seed Name, that document is assigned to all the different entities it refers to.

Regarding the information associated to the entities, for each entity the annotators report (i) the real, anagraphic, name of the entity (based on the annotator knowledge and/or other external resources, (ii) its group name, i.e. the Seed Name, (iii) a free description of the entity, and (iv) any kind of comment it could be necessary.

Another important characteristic of the annotation is the possibility of marking an entity as “similar to” another. This flag is used when the annotator is not sure if two (or more) apparently different entities are the same or not and it can be useful also for evaluation purposes as it allows to change the granularity of the gold standard clustering (more fine-grained if the entities are kept separate or more coarse-grained if they are kept together). All these different kinds of information are annotated into the gold standard through an annotation interface created for that purpose. The interface is described in detail in Section 5.

4. The Gold Standard

The current version of the gold standard is composed as shown in Table 4.

Seed Names	Entities	Documents
209	709	43,704

Table 4. Composition of the gold standard

The next sections report a posteriori data about the actual gold standard corpus after the application of a priori criteria, i.e. (i) the fame of *all* the entities, (ii) the corpus ambiguity of the Seed Names, and (iii) the total number of articles referring to a given Seed Name or a given entity.

4.1. Entity Fame

As already noticed in Section 3.2.1, the entity fame dimension turned out to be difficult to represent, due to the high number of occurrences of famous entities in the corpus. As a matter of fact, the goal of populating each “fame group” with a minimum number of 30 entities each has not been completely reached, as the class of “very famous at the regional level” contains only 24 entities.

Entity fame level	Number of entities
Not famous	542
Quite famous - regional	51
Quite famous - national	44
Very famous - regional	24
Very famous - national	48
Total entities	709

Table 5. composition of the corpus with respect to the entity fame dimension

As expected, we can see in Table 5 that the number of non famous entities is very high in comparison with famous entities. This is due to the fact that given a Seed Name referring to one famous entity, the same Seed Name often refers also to a number of non famous entities.

4.2. Seed Name Ambiguity

The average Seed Name ambiguity in the corpus amounts to 3.39. In order to verify if the PagineBianche can be considered a reliable source for assessing Seed Name ambiguity and if the thresholds we chose are adequate, we calculated the corpus ambiguity of the Seed Names selected from the three ambiguity ranges of PagineBianche. Table 6 shows that there is a correlation between the PagineBianche ambiguity ranges and the actual corpus ambiguity.

PagineBianche ambiguity ranges	Seed Names	Number of Entities	Average corpus ambiguity
Low	121	256	2.12
Medium	42	154	3.67
High	46	299	6.50
All corpus	209	709	3.39

Table 6. Seed Names ambiguity in the corpus

4.3. Number of Documents and Name Variation

The number of documents per entity, after annotation, ranges from 1 document to 1,419, and is well distributed on the whole range.

Among the 32,582 documents containing the Seed Names, 6,637 were not annotated, as they refer to non-person entities or to the journalists who wrote the articles (see Section 3.3).

The total number of documents composing the current version of the gold standard amounts to 43,704, among which 25,945 contain the exact Seed Name and 17,759 contain only name variants.

As regards the different types of name variants occurring in the texts, data about how many name variant types can be found within the annotated documents are not available up to now because the intra-document coreference annotation has not been carried out yet.

5. The Web Annotation Interface

A multi-user web interface was specifically designed for the cross-document coreference annotation task.

The interface is composed of two pages, the *Entity Management Page* and the *Document Annotation Page*, illustrated in Appendix 1. The *Entity Management Page* (Figure 1) contains all information about entities. In the left hand side the *Entity Search* functionality can be found. This functionality allows the annotator to look up a specific entity, to retrieve the list of documents associated to it, and to select the entity for the work session.

In the right hand side of the page, the *Entity Record* and the *Work Session* can be found. The *Entity Record* contains several fields where the annotator inserts and

modifies (i) the real anagraphic name of the entity (e.g. Guido Giuseppe Rossi), (ii) the group name, corresponding to the Seed Name (e.g. Giuseppe Rossi), (iii) a short description characterizing the entity, (iv) the identifier of possible similar entities, (v) an entity fame indicator (according to the annotator’s knowledge), and (vi) a comment with all useful information. Moreover, when necessary, the entity can be marked as “catch all” (see Section 3.3). The *Work Session*, on the bottom right side, contains all the entities created in correspondence with a given Seed Name and is used as entity repository during the document annotation process. In some cases it can happen that two different entities turn out to be the same. The “merge” button allows the annotator to merge the two entities without having to annotate the documents again.

The *Document Annotation Page* (Figure 2) has the same layout of the OntoText Portal. The annotator submits a query and obtains all the documents satisfying the query, together with the text snippet in which the query string occurs.

A scroll down menu is associated to each retrieved document, where the annotator can select the entity to which the document refers. The entities presented for annotation correspond to those inserted in the Work Session created by the annotator in the Entity Management Page. If the document snippets are not informative enough to individuate the correct entity, the annotator can also access the full article. If the document turns out to be really difficult to be assigned to an entity, the annotator can mark the annotation as “not sure” by clicking the button at the left of the scroll down menu.

When the results of an annotator query are displayed, all the documents already annotated according to the entities contained in the Work Session are highlighted and the annotator can decide if he/she prefers to see them in the page or to hide them.

6. Conclusion and Future Work

We presented work aimed at developing a gold standard for person cross-document coreference resolution. The first version contains 209 different names, 709 different entities, and more than 43,700 newspaper articles.

We think that such an extensive gold standard can help assess and advance the state of the art for cross-document entity coreference resolution. However, the sampling criteria followed to generate the gold standard, especially the suitability of the external source used to determine Seed Name ambiguity, the method of evaluation of the entity fame, and the balancing of these two dimensions, represent issues which are open to discussion.

Up to now, we have gathered the data necessary to calculate the inter-annotator agreement, which will refer to 20 Seed Names (10% of the total) selected from the different cells composing the gold standard. The annotation has been performed by two of the five annotators who worked at the gold standard.

As regards the metrics to be used to calculate intercoder agreement, different measures have been proposed in the literature in the last years, the most used for NLP tasks being the K measure. Recently, the suitability of the traditional K measure has been put under discussion (Artstein and Poesio, to appear). As regards our specific field, the main problems relate to the fact that (i) in a

clustering task there is not a common and predefined set of categories (the different person entities), and (ii) the distribution of the number of clusters and their size is not homogeneous among the different Seed Names. We have not calculated inter-annotator agreement yet. However, as a preliminary assessment, we carried out the evaluation of the two manual annotations with the SemEval-2007 Web People Search scorer. The scorer relies on the standard clustering measures of Purity, Inverse Purity, and F-measure. Table 7 reports the results obtained for the annotators, which can be compared with the “All-in-One” baseline run on Adige-500K.

	Purity	Inverse Purity	F-measure
Annotators	0.92	0.90	0.91
Baseline	0.86	0.77	0.81

Table 7. Preliminary evaluation of inter-annotator agreement

Table 7 shows that the manual annotation outperforms the All-in One baseline, suggesting that our gold standard has been annotated with a good intercoder agreement. Annotator 1 created 88 entities and annotated 5,176 documents, while Annotator 2 created 103 entities and annotated 5,030 documents.

As further future work, we plan to carry out the annotation of the intra-document coreference using the name variants found during the cross-document annotation.

Both the design of the gold standard and the various kinds of information contained in the annotations allow a wide range of possible evaluations.

The partition of the gold standard in 15 classes, representing the different levels of entity fame and Seed Name ambiguity, allows for a more informative evaluation and analysis of systems performances.

Concerning the task to be evaluated, exact name and name variations are considered, thus covering the whole cross-document coreference spectrum. As regards the evaluation itself, it is possible to set the gold standard clustering granularity (grouping or maintaining separate entities marked as similar) and to assign different scores to documents marked as “not sure” for the cluster to which they have been linked.

The first usage of the corpus has been the evaluation of the OntoText coreference algorithm (Popescu and Magnini 2007, Popescu 2008). To this purpose, we exploited the SemEval scorer.

As regards other uses of the gold standard, when the intra-document coreference annotation will be performed, it will also be possible to evaluate this task. Finally, we envisage its usage within the next edition of EVALITA (Magnini and Cappelli, 2007), a new initiative devoted to the evaluation of Natural Language Processing tools for Italian.

7. Acknowledgements

This work has been supported by the Ontotext project, funded by the Autonomous Province of Trento under the FUP-2004 research program. Our sincere thanks go to Diego Barbon, Giampier Bondoni, Fabio Niccolini,

Giuliano Tomasini, and Matteo Tonini who contributed to the development of the gold standard.

8. References

- ACE 2008 Evaluation Plan
<http://www.nist.gov/speech/tests/ace/2008/doc/ace08-evalplan.v1.1.pdf>
- Artiles, J., Gonzalo, J., Sekine, S. (2007). Establishing a benchmark for the Web People Search Task: The Semeval 2007 WePS Track. In *Proceedings of SemEval-2007 Workshop*, co-located with ACL 2007, Prague, CZ, 23-24 June 2007.
- Artstein, R., Poesio, M. (to appear). Inter-Coder Agreement for Computational Linguistics. In *Computational Linguistics*, to appear.
- Blume, M. (2005). Automatic Entity Disambiguation: Benefits to NER, Relation Extraction, Link Analysis, and Inference. In *Proceedings of the International Conference on Intelligence Analysis*, McLean, Virginia, USA, 2005.
- Bollegala, D., Matsuo, Y., Ishizuka, M. (2006). Extracting Key Phrases to Disambiguate Personal Name Queries in Web Search. In *Proceedings of the ACL'06 Workshop on How Can Computational Linguistics Improve Information Retrieval?*
- Giuliano, C. (2002) A Tool-Box for Lexicographer. In *Proceedings of the Tenth EURALEX Congress*, Denmark, Copenhagen, August 13-17, 2002.
- Gooi, C. H., Allan, J. (2004). Cross-Document Coreference on a Large Scale Corpus. In *Proceedings of the Human Language Technology Conference (HLT-NAACL)*. Boston, Massachusetts, 2004.
- Magnini, B., Cappelli, A. (2007). Evalita 2007: Evaluating Natural Language Tools for Italian. Special issue of *Intelligenza Artificiale (AI*IA)*.
- Magnini, B., Pianta, E., Girardi, C., Negri, M., Romano, L., Speranza, M., Bartalesi-Lenzi, V., Sprugnoli, R. (2006). I-CAB: the Italian Content Annotation Bank. In *Proceedings of LREC 2006 - Genova, Italy*, May 22-28, 2006.
- Mann, G. S., Yarowsky, D. (2003). Unsupervised personal name disambiguation. In *Proceedings of CoNLL-7*, 2003.
- Popescu, O., Magnini, B. (2007). Web People Search Using Name Entities. In *Proceedings of SemEval-2007 Workshop*, co-located with ACL 2007, Prague, CZ, 23-24 June 2007.
- Popescu, O. (2008). *Ontological Constraints in a Statistical Framework for Person Cross Document Coreference*. PhD Thesis, University of Trento. FBK Technical Report, April 2008

Appendix 1: the Annotation Interface

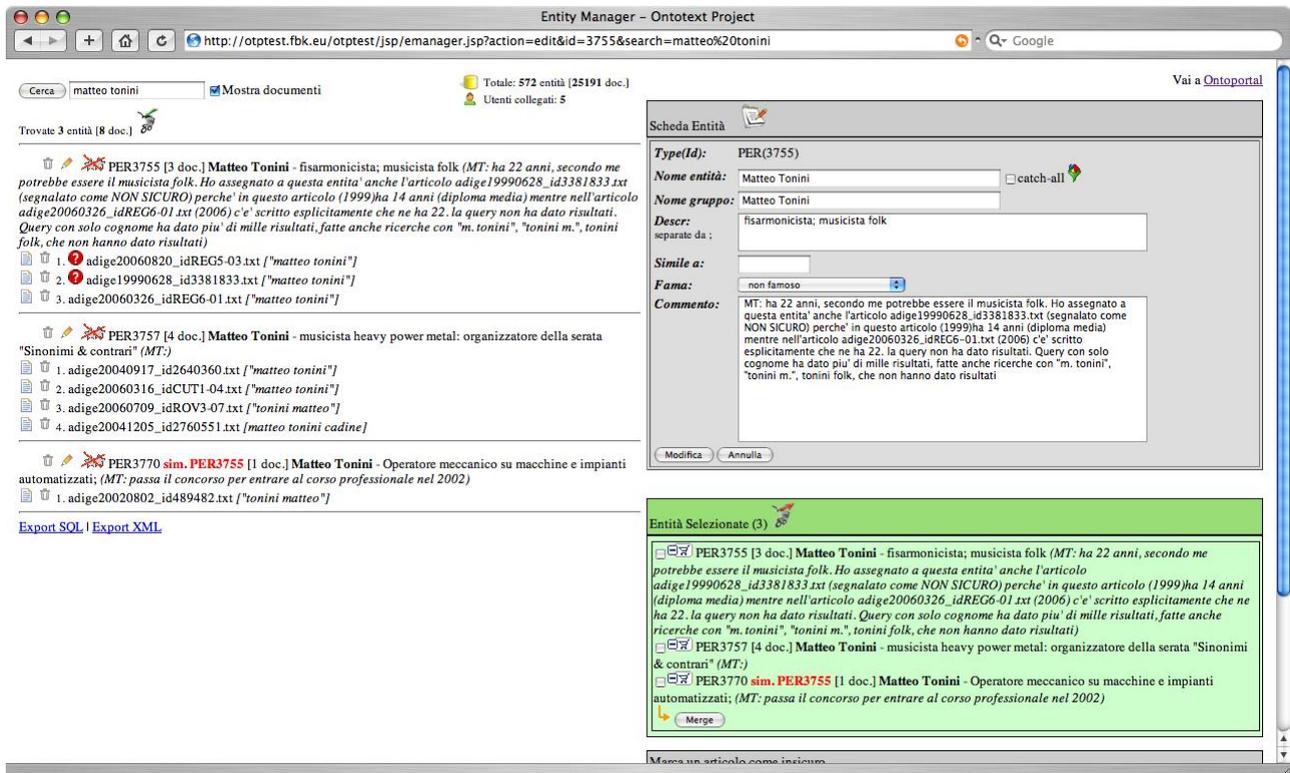


Figure 1. The Entity Management Page

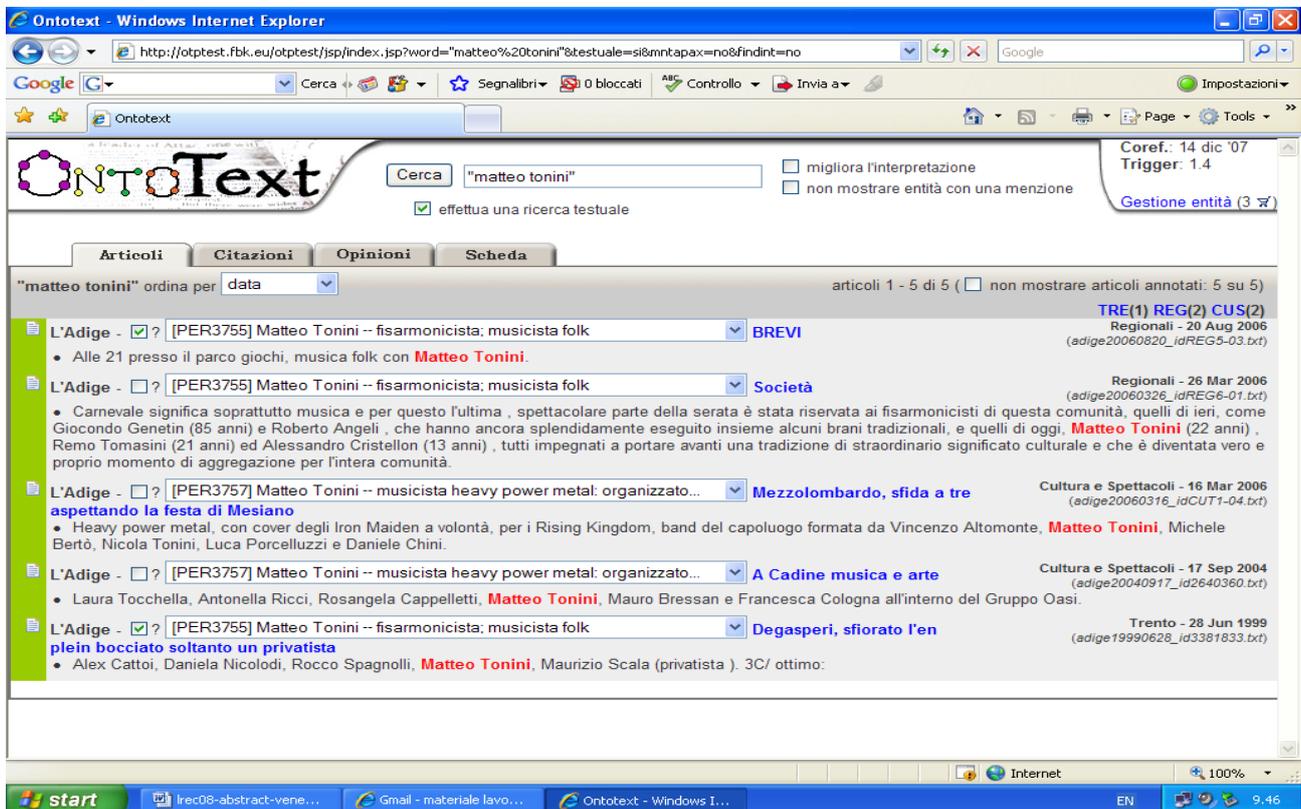


Figure 2. The Document Annotation Page