Applying Natural Language Processing to Speech Transcriptions for Automated Analysis of Educational Video Broadcasts

Maurizio Montagnuolo, Roberto Borgotallo, Silvia Proscia, Laurent Boch

RAI Radiotelevisione Italiana
Centro Ricerche e Innovazione Tecnologica, Torino, Italy
{maurizio.montagnuolo, roberto.borgotallo, silvia.proscia, laurent.boch}@rai.it

Abstract. This paper describes the results of a work, carried out by RAI within the framework of the project "La Città Educante", aiming at creating statistical models for automatic document categorization and named entity recognition, both acting in the educational field and in Italian language. The taxonomy used for the documents categorization is the Scientific Disciplinary Sector taxonomy (SSD) used in Italy to organize the disciplines and thematic areas of higher education. The actual statistical models were created with the Apache OpenNLP libraries. The obtained results showed fairly good accuracy with SSD document classification and some significant improvement of categorization precision for Named Entities compared to the state of the art.

Keywords: video categorization, ontology, named entity recognition

1 Introduction and Related Work

With the increase of the number of multimedia contents that are produced everyday by a wide range of applications and devices, users can now access huge amount of data in wide-world localized archives and repositories. As an example, in the educational domain, teachers can now circulate their lectures through online video repositories, like e.g. VideoLectures.NET.[1] Effective and comprehensive annotations are needed in order to quickly and easily access those contents. In the literature many approaches can be found that aim at addressing the problem of video analysis and retrieval. Thought, almost all of the proposed research targets specific and restricted domains such as sports and news broadcasting [8, 9, 13, 14, 18], and very few that specifically targets the educational domain. In [7] a visual content classification system (VCCS) combining support vector machine (SVM) and optical character recognition (OCR) to classify visual content into figures, text and equations was proposed. Basing on the assumption that text in lecture video is directly related to the lecture content, in [17] an approach for automated lecture video indexing based on video OCR technology

[1] http://videolectures.net/ (last accessed September 2017)

was presented. Speech is the most natural way of communication carrying most of information in nearly all lectures. Therefore, it is of clear advantage that the speech information can be used for automated analysis and annotation of lecture videos. In [6] a proposal to transform the video subject classification problem into a text categorization problem by exploiting the extracted transcripts of videos was presented. In [16] both video OCR and automatic speech recognition (ASR) were applied to develop a framework for automated analysis and indexing of German video lectures. When a large number of videos is collected, filtering them by subject is especially important because it allows users (e.g. students) to find specific information on desired themes. However, subjects may often overlap to each other, such as for example medicine and biology or telecommunications and computer science. To solve this problem, fuzzy clustering for video lecture classification was proposed in [4].

"La Città Educante" project[2] aims to develop new models of learning and teaching by exploring, developing and evaluating innovative technologies for knowledge extraction, management and sharing. These include content based video analysis, segmentation and summarization [1, 2, 10], text analysis and semantic categorization, data mashups and visualization. This paper describes the use of natural language processing for the development of a framework for automatic annotation of educational video broadcasts. The problem is switched from the video domain to the textual domain by performing ASR conversion on the input video and then applying document categorization and named entity recognition on the transcribed texts. The remaining of the paper is organized as follows. Section 2 describes the theoretical background on which this work is grounded. Section 3 describes the experimental study we conducted. Section 4 concludes the paper with some summary remarks.

2 Video Annotation as Text Analysis Process

This section describes the theoretical background on which automatic video annotation was performed. The assumption is that video annotation can be shifted to a language processing problem by analyzing the text automatically derived by the speech content of broadcast material. Two problems are approached, i.e. document categorization and named entity recognition from spoken documents. To address these problems, the Apache OpenNLP library[3] was used.

2.1 Categorization Criteria

RAI has a multi-year experience in automatic metadata extraction in the news domain and developed technologies used internally also for automatic news categorization. In this context a proprietary classification schema already adopted in manual documentation is used [11]. The educational context of "La Città Educante", despite some similarities, is clearly different. Therefore, we decided to

[2] `http://www.cittaeducante.it` (last accessed September 2017)
[3] `https://opennlp.apache.org/` (last accessed September 2017)

adopt a formal and authoritative classification schema conveying the subjects of study of secondary (lower and/or higher) and university schools. The following subsections describe the classification system used and its implementation in accordance with the Semantic Web techniques.

Definition of the Categorization System. In order to define the most appropriate classification schema various options were considered, including the names of the subjects of study in various degrees of education. As an alternative, the organization of the contents of Skuola.net,[4] a portal that keeps and shares teaching materials, notes and news addressed to both middle/high school and university students, was considered. Starting from this analysis, it can be observed that topics are more and more specific as far as the level of education increases. However grouping in areas does not follow a coherent set of criteria, as it may be appropriate to reflect the goals of the different institutes. In the case of Skuola.net, topics are mapped to the specific faculties and examinations that meet the greatest interest of the portal users. This latter approach is certainly interesting, as it represents a categorization that takes into account the point of view of the final user. Though, in the absence of a specific model, it would be preferable to adopt a set of neutral categories with respect to the goals of particular communication operators. Thus, we decided to adopt the Scientific Disciplinary Sectors (SSD) classification system,[5] which is used in Italy for the organization of higher education as the reference for the subject classification. The SSD, defined by the Italian Ministry of Education, University and Research (MIUR), is organized hierarchically on four levels. There are 14 areas at the top level, 88 macro-sectors at the second level, 188 competition sectors at the third, and 367 scientific-disciplinary sectors at the last level. Hierarchical classification allows the attribution of one (or more) category(s) to an appropriate degree of specificity/generality. We limited the work to the first two levels of the hierarchy, as the number of remaining classes is comparable to what already done for journalism in RAI [12]. Nothing prevents, however, to apply subsequently a dynamic criterion so that particularly populous classes can be further subdivided into sub-categories. The OpenNLP Document Categorizer tool was used to perform automatic categorization of input texts according to the defined taxonomy. This tool uses the maximum entropy framework to train a categorization model based on pre-annotated corpora.

Categorization System Implementation. Semantic Web aims to foster knowledge sharing through interoperability and data exchange. For this purpose, the World Wide Web Consortium (W3C) defined a number of formal languages that allow the development of systems for the integration of heterogeneous data sources, such as those from multimedia archives, social networks and open data.

[4] `http://www.skuola.net/` (last accessed September 2017)

[5] `http://attiministeriali.miur.it/anno-2015/ottobre/dm-30102015.aspx` (last accessed September 2017)

Among them, the Simple Knowledge Organizational System (SKOS)[6] represents the state of the art for describing structured vocabularies, such as glossaries, classification systems, and taxonomies. SKOS was designed with the idea of being easily extensible in order to allow connection and sharing of knowledge between different databases. The main SKOS elements are:

- `Concepts`, which are elementary information units, representing ideas, meanings, objects, or events at the basis of a system of knowledge organization. Concepts can be grouped into collections and/or organized according to a particular schema (`ConceptScheme`);
- `Labels` and `Notations`, which are a set of words or phrases, even in different languages, that describe a concept;
- Relationships, such as hierarchies and equivalences, between concepts within the same dictionary (`Semantic Relations`) and between concepts defined in different dictionaries (`Mapping Relations`).

Some examples of taxonomies implemented in SKOS include the multilingual thesaurus of the European Union (Eurovoc),[7] and the thesaurus of the Wikipedia categories. As part of "La Città Educante", the Scientific Disciplinary Sectors (SSD) taxonomy has been implemented in SKOS through an iterative process defined by the following steps:

- **Definition of concepts.** A SKOS concept was created for each SSD area and macro-sector and identified by the corresponding alphanumeric code defined by the ministerial ordinance. For example the code 12/F indicates the macro-sector F in area 12 of the SSD;
- **Definition of the main scheme.** In order to group all the concepts of the ontology, a top level SKOS `ConceptScheme` was defined to which all the SSD areas are related by the `hasTopConcept/topConceptOf` property of the SKOS standard;
- **Concept description.** Each concept was described with a label (`prefLabel`) based on the name (Italian and English) of the corresponding denomination defined by the ministerial decree. For example, the '*Diritto processuale civile*' and '*Civil procedural law*' labels were assigned to the macro-sector 12/F, respectively for the Italian and English languages;
- **Definition of intra-schema hierarchy.** Hierarchical relationships between areas and macro-sectors were defined using the `broader` and `narrower` properties of the SKOS standard;
- **Definition of inter-schema relationships.** In order to promote interoperability with other classification schemes, each SSD concept was mapped with at least one concept of the Eurovoc thesaurus and of DBpedia categories. Eurovoc was selected as it allows a unique classification of documents in the European Union's institutional databases irrespective of the language used in the documents themselves. DBpedia was chosen as the central hub

[6] `https://www.w3.org/2004/02/skos/` (last accessed September 2017)
[7] `http://eurovoc.europa.eu/` (last accessed September 2017)

```
<!-- http://attiministeriali.miur.it/anno-2015/ottobre/dm-30102015#06 -->

<owl:NamedIndividual rdf:about="http://attiministeriali.miur.it/anno-2015/ottobre/dm-30102015#06">
  <rdf:type rdf:resource="&skos;Concept"/>
  <skos:notation rdf:datatype="&rdfs;Literal">06</skos:notation>
  <skos:prefLabel xml:lang="en">MEDICINE</skos:prefLabel>
  <skos:prefLabel xml:lang="it">SCIENZE MEDICHE</skos:prefLabel>
  <skos:narrower rdf:resource="http://attiministeriali.miur.it/anno-2015/ottobre/dm-30102015#06/A"/>
  <skos:narrower rdf:resource="http://attiministeriali.miur.it/anno-2015/ottobre/dm-30102015#06/B"/>
  <skos:narrower rdf:resource="http://attiministeriali.miur.it/anno-2015/ottobre/dm-30102015#06/C"/>
  <skos:narrower rdf:resource="http://attiministeriali.miur.it/anno-2015/ottobre/dm-30102015#06/D"/>
  <skos:narrower rdf:resource="http://attiministeriali.miur.it/anno-2015/ottobre/dm-30102015#06/E"/>
  <skos:narrower rdf:resource="http://attiministeriali.miur.it/anno-2015/ottobre/dm-30102015#06/F"/>
  <skos:narrower rdf:resource="http://attiministeriali.miur.it/anno-2015/ottobre/dm-30102015#06/G"/>
  <skos:narrower rdf:resource="http://attiministeriali.miur.it/anno-2015/ottobre/dm-30102015#06/H"/>
  <skos:narrower rdf:resource="http://attiministeriali.miur.it/anno-2015/ottobre/dm-30102015#06/I"/>
  <skos:narrower rdf:resource="http://attiministeriali.miur.it/anno-2015/ottobre/dm-30102015#06/L"/>
  <skos:narrower rdf:resource="http://attiministeriali.miur.it/anno-2015/ottobre/dm-30102015#06/M"/>
  <skos:narrower rdf:resource="http://attiministeriali.miur.it/anno-2015/ottobre/dm-30102015#06/N"/>
  <skos:topConceptOf rdf:resource="http://attiministeriali.miur.it/anno-2015/ottobre/dm-30102015#SSD"/>
  <skos:exactMatch rdf:resource="&dbpedia;/Category:Medicine"/>
  <skos:exactMatch rdf:resource="&eurovoc;/5881"/>
  <skos:exactMatch rdf:resource="&nubilarium;medicina-e-salute"/>
</owl:NamedIndividual>
```

Fig. 1. Excerpt from the SSD ontology.

of the Linked Open Data (LOD) network. Furthermore, in order to ensure interoperability with the RAI archives, each SSD concept was mapped to the classification schema used for documentation of the RAI's archives.

Inter-schema relationships were defined according to the following criteria:

- The SKOS `hasExactMatch` relates concepts identified by the same label to the considered ontologies. For example, the SSD macro-sector ('*Civil Procedure Law*') was mapped exactly to the '*Civil_procedure*' DBpedia category;
- The SKOS `hasNarrowMatch` property (`hasBroadMatch`) relates more in-depth (generic) concepts to the considered ontologies. For example, the SSD macro-sector '*International Law, European Union, Comparison, Economy, Markets and Navigation*' `hasNarrowMatch` '*Business_law*', '*Comparative_law*' , '*European_Union_law*', and '*International_law*' by DBpedia;
- The SKOS `hasRelatedMatch` property relates concepts that share one or more features but whose mapping do not fall into any of the previous cases. For example, the SSD macro-sector '*Astronomy, Astrophysics, Physics of Earth and Planets*' `hasRelatedMatch` '*Planetary_science*' by DBpedia.

An excerpt from the ontology is shown in Fig.1.

2.2 Named Entity Recognition

For the creation of the named entity recognition (NER) models we adopted a semi-supervised approach, a technique commonly used in machine learning when,

given a large dataset, only a subset has annotations. The complete manual annotation of the entire dataset is a long and expensive process not exempt from human mistakes. E.g., Xue et al [15] reported that approximately two years of work were needed for the manual annotation of about 4,000 sentences for natural language analysis applications. The actual creation and assessment of the NER models was done in three steps. First, a set of web articles falling into the project's educational scope (talking about economics, science, technology, health, etc.) was selected among a larger and more generic set (including also off-topics like sport, talking about politics and others) of news articles pre-annotated with named entities information (persons, locations and organizations) by an automatic system already in possession of RAI. These selected texts were adapted to make them similar to those coming out of a basic ASR process, by removing punctuation and capital letters in order to create the final training corpora resulted in a set of around 47,000 sentences. This number is three times larger than the minimum size recommended by the Apache OpenNLP documentation. In the second step, we used the `TokeNameFinderTrainer` tool in the Apache OpenNLP library to generate the new NER models. Similarly to document categorization, the `TokeNameFinderTrainer` tool creates a maximum-entropy-based name finder, hence the output of this step consists of three binary files representing the OpenNLP models for the corresponding categories of entities considered. At last, we run the new models on a set of automatic speech transcriptions from RAI's broadcasts and manually validated them by assessing the entities found within the input material.

3 Experimental Results

This section describes the experiments that we undertook to demonstrate the effectiveness of the presented methods. The reference dataset includes 2,020 episodes of TGR Leonardo, a science newscast produced by RAI and broadcasted daily from Monday to Friday on the RAI3 channel. Each episode, lasting approximately 10 minutes, features news about technology, health, economy, environment and society. The format is that of a traditional newscast. Each episode was automatically segmented and transcribed into elementary news stories using the RAI ANTS system [11], resulting in about 6,600 news items.

3.1 Ground Truth Generation for Document Categorization

This section describes the guidelines adopted at the annotation stage as well as the composition and organization of the reference dataset. Assigning SSD areas and macro-sectors requires a good understanding of the hierarchy of categories and a careful listening and evaluation of content. The hierarchy of categories needs to be well-known as some words can have wider or narrower meaning with respect of common usage. For example the term pedagogy does not simply indicate the educational aspect of children/adolescents (as often intended) but of people of all ages. To create the ground-truth categories we assigned one or

two as a maximum areas/macro-sectors to each news item, according to their importance with respect to the news item topic. For example, if a news item talks about the use of information technology in secondary schools, the news item would have been annotated by '*History, philosophy, pedagogy and psychology*' and '*Mathematics and informatics*' as, respectively, primary and secondary SSD area. All detected news items were manually annotated by a group of 10 people. The total number of annotated news is 6,608, corresponding to approximately 243 hours of audio-visual material. The size of this dataset is in line with those adopted in similar works [5]. The distribution of the annotated SSD areas is shown in Fig.2. The number of news annotated with two areas is 1,593 that corresponds around a quarter of the total (i.e. about 6,600). Considering only the main category (Area1) there is a fairly good coverage of the annotations. For the purpose of training an automatic classification system, the use of a second category does not add significant information. However, it is interesting to analyze the relationships between the areas in multiple annotation cases, as shown in Fig.3. From these data the following cases of interest can be identified:

- **Mono-area and multi-sector news items.** A news item within the same area that may be classified according to more than one macro-sector within the same area;
- **Multidisciplinary news items.** A news item that may be classified in two or more different areas. For example, a news item on legal regulations on the use of drugs might be annotated primarily as *Juridical Sciences* (Area 14) and secondly as *Biological Sciences* (Area 5);
- **Doubtful news items.** A news item that, due to the nature of its topic, may be misclassified. For example, the area of *Biological Sciences* (Area 5) might be confused with the area of *Medical Sciences* (Area 6) and vice versa.

Similarly, we evaluated the distribution of SSD macro-sectors compared to annotated news. The number of annotations to get a uniform coverage of all the macro-sectors is 100. All macro-sectors were annotated at least once. 25% of them were annotated for a sufficiently representative number of times while 50% of them have less than half of the optimum number of annotations.

3.2 Document Categorization Performance Evaluation

A 10-fold cross validation was used to assess the performance of the categorization task. Fig.4 shows the results on the average precision, recall, and F-measure for the classification according the SSD areas. The categorization model achieves good performance in terms of precision.

Analyzing average recall values, a more varied situation is noted. In particular, for the areas most represented in the reference dataset, a good balance between precision and recall is obtained. Conversely, for the areas that are least represented in the dataset, low recall values are obtained. This may depend on the use of a not enough large learning dataset (e.g. insufficient annotations for a given area), or on the interchangeability of the area with other areas with

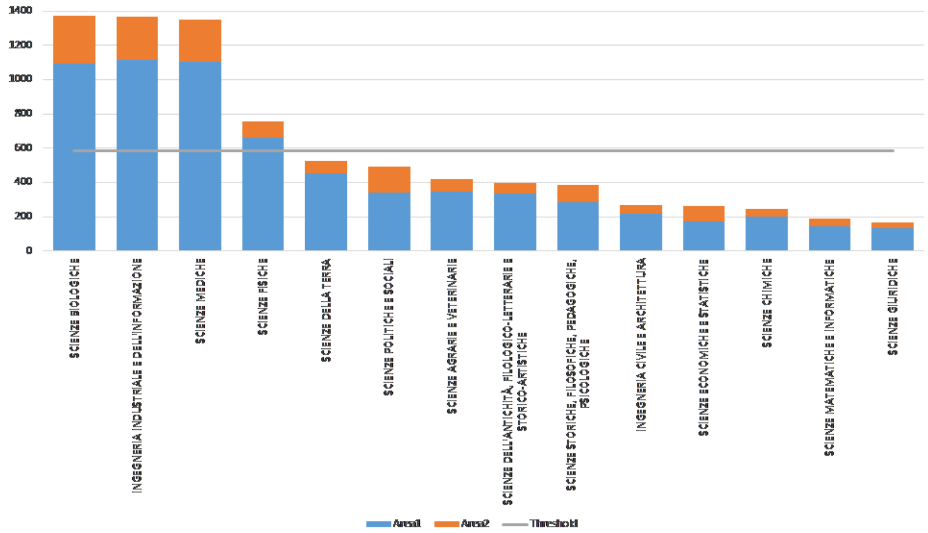

Fig. 2. Distribution of SSD areas with respect to annotated audiovisual clips.

similar topics. This phenomenon is confirmed by the analysis of the confusion matrix shown in Fig.5. The matrix rows show the expected areas. The matrix columns count the deduced areas (i.e. those generated automatically by the categorizer). The element in position (i,j) indicates the percentage of news items belonging to area i that have been categorized with the area j. The matrix diagonal shows the percentage of news items ranked correctly for each area, while the extra-diagonal elements contain misclassification errors. It can be noted that the matrix has the maximum at the diagonal for most of the considered areas (highlighted in green). Areas subject to many categorization errors are highlighted in red. For these areas, errors are mostly distributed along the matrix columns, confirming good precision with respect to lower recall. Some areas tend to be confused more than others (e.g., *Medical Sciences* with *Biological Sciences* and vice versa), in accordance with intrinsic ambiguity (subject matter) or derivative (annotation errors) of the same topics.

Similarly, we evaluated the performance of the categorization of the SSD macro-sectors. Table 1 reports the top-5 classification scores. As expected, global performance is worse than performance of the SSD area categorizer. This might be attributable to different causes, including the small size of training examples for some macro-sectors, the higher number of categories (i.e. macro-sectors) to be recognized, and the higher degree of interchangeability among them.

3.3 Named Entity Recognition Performance Evaluation

NER process was evaluated in terms of precision for each of the entity categories considered. Recall was omitted from the analysis because of the huge amount

P(Area1 = i, Area2 = j)		Area 2														
		1	2	3	4	5	6	7	8	9	10	11	12	13	14	P(Area1 = i)
Area 1	1	0.19%	0.13%	0.06%	0.06%	0.13%	0.31%	0.00%	0.06%	1.44%	0.38%	0.19%	0.00%	0.31%	0.69%	3.95%
	2	0.19%	1.51%	0.25%	0.69%	0.50%	0.06%	0.06%	0.06%	3.58%	0.19%	0.19%	0.00%	0.06%	0.50%	7.85%
	3	0.00%	0.38%	0.19%	0.06%	0.63%	0.82%	0.13%	0.06%	0.31%	0.00%	0.06%	0.25%	0.19%	0.38%	3.45%
	4	0.13%	0.75%	0.13%	0.00%	1.69%	0.13%	0.06%	0.82%	0.69%	0.25%	0.56%	0.13%	0.44%	0.38%	6.15%
	5	0.19%	0.31%	0.56%	0.94%	5.27%	3.89%	1.95%	0.13%	0.75%	0.88%	0.82%	0.31%	0.69%	1.95%	18.64%
	6	0.38%	0.06%	0.82%	0.13%	3.64%	6.40%	0.88%	0.06%	1.13%	0.31%	1.00%	0.19%	0.56%	1.76%	17.33%
	7	0.00%	0.00%	0.25%	0.06%	0.69%	0.38%	0.88%	0.19%	0.25%	0.00%	0.06%	0.44%	0.13%	0.13%	3.45%
	8	0.00%	0.06%	0.19%	0.44%	0.31%	0.19%	0.00%	0.44%	0.63%	0.38%	0.06%	0.00%	0.19%	0.06%	2.95%
	9	0.50%	2.13%	0.19%	1.26%	1.76%	0.69%	0.06%	0.75%	4.27%	0.38%	0.82%	0.38%	0.75%	0.94%	14.88%
	10	0.19%	0.06%	0.00%	0.19%	0.38%	0.13%	0.06%	0.25%	0.82%	0.88%	0.75%	0.06%	0.00%	0.44%	4.21%
	11	0.44%	0.31%	0.06%	0.19%	0.50%	0.94%	0.00%	0.13%	0.94%	0.06%	0.56%	0.19%	0.19%	1.13%	5.65%
	12	0.13%	0.06%	0.06%	0.06%	0.69%	0.31%	0.19%	0.06%	0.50%	0.06%	0.06%	0.00%	0.44%	0.25%	2.89%
	13	0.31%	0.00%	0.13%	0.00%	0.38%	0.25%	0.19%	0.19%	0.25%	0.06%	0.13%	0.19%	1.07%	0.75%	3.89%
	14	0.19%	0.31%	0.00%	0.31%	0.88%	0.75%	0.06%	0.00%	0.25%	0.06%	0.82%	0.13%	0.63%	0.31%	4.71%
	P(Area2 = j)	2.82%	6.09%	2.89%	4.39%	17.45%	15.25%	4.52%	3.20%	15.82%	3.89%	6.09%	2.26%	5.65%	9.67%	100.00%

Fig. 3. Co-occurrence (Area1, Area2) in the annotation dataset. In yellow, the most frequent combinations are highlighted.

Table 1. Top-5 Precision, Recall and F-measure for macro-sector categorization.

Macro-sector	**Precision**	**Recall**	**F-measure**
Astronomia, astrofisica, fisica della terra e dei pianeti	0.55	0.67	0.61
Ingegneria energetica, termo-meccanica e nucleare	0.44	0.58	0.50
Scienze archeologiche	0.49	0.47	0.48
Ingegneria meccanica, aerospaziale e navale	0.39	0.61	0.47
Geoscienze	0.37	0.59	0.46

of time required to assess it (since it requires having a dataset fully annotated). This omission is considered acceptable because in our application scenario the Recall is less significant. When doing searches on large amount of data in fact, users usually obtain a lot of matching documents and desire that they are not false positives in order not to waste their time, this means that high precision is desirable. For sure low recall excludes from the search potentially interesting documents but this is less important for not really specific searches made on big indexes. The dataset of 6,600 TG Leonardo news items was randomly split into four sets of equal size, each of them assigned to one annotator for evaluation. Precision was assessed manually, counting the occurrences of the following cases:

- **Correct identification and classification.** Entity correctly identified in the text and attributed to the right category.
- **Correct identification but incorrect classification.** Entity correctly identified in the text but attributed to a wrong category.
- **Wrong identification.** Entity mistakenly identified.

The results obtained are shown in Table 2. For calculation purpose, the Named Entities found were considered as statistically independent although present with multiple occurrences in the same document, e.g. the word "Italia" found twice as a location in the same text counted as two correct occurrences and not just one. This method is justified by the fact that identification and classification are carried out on the basis of context analysis limited to the surrounding words and not with simple match on a vocabulary. As a first consideration of data analysis, it is noted that the obtained precision values are aligned between the four

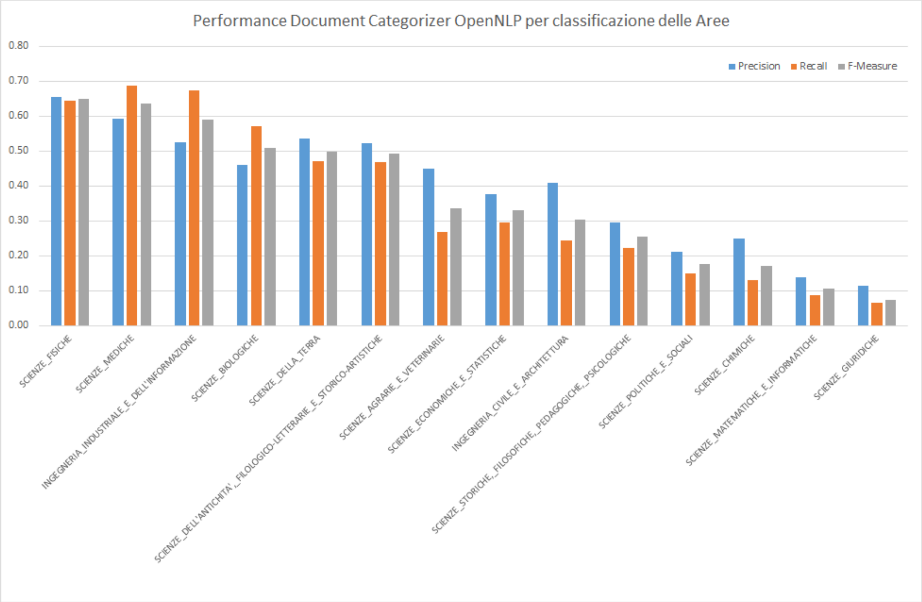

Fig. 4. SSD area classification model performance measurement.

Media - Confusion Matrix (%)	INGEGNEI	SCIENZ	SCIENZE_BI	SCIENZE_F	SCIENZ	SCIENZE_AC	SCIENZE_F	SCIENZ	SCIENZE_ST	INGEGNEI	SCIENZ	SCIENZE_EC	SCIENZE_F	SCIENZE_GIURIDICHE
INGEGNERIA_INDUSTRIALE_E_DELL'INFORMAZIONE (9)	67	4	4	8	3	0	2	2	3	2	0	1	2	1
SCIENZE_MEDICHE (6)	3	69	15	0	0	2	3	1	3	0	2	1	1	0
SCIENZE_BIOLOGICHE (5)	5	18	57	1	4	4	3	2	1	1	1	1	0	1
SCIENZE_FISICHE (2)	22	1	4	65	4	0	1	0	1	0	2	0	1	0
SCIENZE_DELLA_TERRA (4)	11	2	21	7	47	0	3	3	2	3	0	1	1	1
SCIENZE_AGRARIE_E_VETERINARIE (7)	6	10	47	1	1	27	1	0	1	1	2	2	0	3
SCIENZE_POLITICHE_E_SOCIALI (14)	14	18	16	4	4	2	13	3	10	1	2	6	3	2
SCIENZE_DELL'ANTICHITA'_FILOLOGICO-LETTERARIE_E_STORICO-ARTISTICHE (10)	15	3	14	4	2	1	4	47	5	5	0	0	3	0
SCIENZE_STORICHE,_FILOSOFICHE,_PEDAGOGICHE,_PSICOLOGICHE (11)	15	17	10	4	3	0	13	9	22	1	0	3	3	1
INGEGNERIA_CIVILE_E_ARCHITETTURA (8)	36	4	9	1	10	1	1	8	1	28	2	2	0	1
SCIENZE_CHIMICHE (3)	11	23	25	8	2	8	3	2	1	2	13	2	1	1
SCIENZE_ECONOMICHE_E_STATISTICHE (13)	17	10	13	0	5	4	10	1	2	2	1	30	2	3
SCIENZE_MATEMATICHE_E_INFORMATICHE (1)	42	10	1	5	2	0	8	4	8	0	2	8	9	1
SCIENZE_GIURIDICHE (12)	23	6	23	3	2	6	15	1	4	1	5	4	2	7

Fig. 5. Confusion matrix for classifying SSD areas.

datasets, confirming a substantial validity and concordance of assessments over the four annotators. The categorization precision values are over 70% (i.e. more than double a random grading process characterized by a uniform probability distribution) for each of the test datasets. These are therefore good results, even more significant when compared with the state of the art [3], where maximum precision of 65% is reported.

4 Conclusions

This paper described the results of a work carried out by RAI, aimed to create statistical models for automatic document categorization and named entity recognition, both acting in the educational field and in Italian language. The taxonomy used for the documents categorization is the Scientific Disciplinary Sector

Table 2. Named Entity Recognition performance. LOC (Locations), PER (Persons), ORG (Organizations). For each of the four subsets of documents considered and for each of the named entity category are reported: Precision of the detection process; Precision of the classification process (Category Precision); Global Precision.

	Detection Precision	**Category Precision**	**Global Precision**
Dataset 1			
LOC	91.54	98.36	89.9
PER	77.08	98.58	75.66
ORG	72.25	98.20	70.45
Dataset 2			
LOC	91.07	99.75	90.82
PER	71.58	100	71.58
ORG	75.90	97.95	73.85
Dataset 3			
LOC	88.90	98.89	87.78
PER	70.98	98.85	69.83
ORG	60.22	98.90	59.12
Dataset 4			
LOC	91.44	99.38	90.82
PER	73.10	99.63	72.73
ORG	71.61	96.45	68.06

taxonomy that, as part of the work, has been coded using SKOS to get a formal XML/OWL ontology. The reference ground-truth for document categorization has been created by annotating a dataset of automatic speech transcriptions derived from RAI's scientific newscasts. Named entity recognition models were generated to be suitable for automatically transcribed text without punctuation and capital letters. Document categorization performance was calculated in terms of precision, recall and F-measure. Named entity recognition was evaluated in terms of precision of both entity detection and entity classification, distinctly for locations, people and organizations. The obtained results showed fairly good accuracy with SSD document classification and some significant improvement of categorization precision for Named Entities compared to the state of the art. Next step for RAI is to test the trained OpanNLP models both in the context of the "La Città Educante" and for an internal archive search and retrieval experimental service. As an example, locations automatically extracted within speech transcriptions could be used to enrich archive metadata and provide auxiliary information for applications such as georeferencing, data mashups and map visualization.

Acknowledgments. This work was carried out within the project "La Città Educante" (CTN01 00034 393801) of the National Technological Cluster on Smart Communities cofunded by the Italian Ministry of Education, University and Research - MIUR.

References

1. Baraldi, L., Grana, C., Cucchiara, R.: Recognizing and presenting the storytelling video structure with deep multimodal networks. Trans. Multi. 19(5), 955–968 (2017)
2. Baraldi, L., Grana, C., Messina, A., Cucchiara, R.: A browsing and retrieval system for broadcast videos using scene detection and automatic annotation. In: Proc. of the 2016 ACM on Multimedia Conference. pp. 733–734. MM '16 (2016)
3. Bartalesi Lenzi, V., Speranza, M., Sprugnoli, R.: Named Entity Recognition on Transcribed Broadcast News at EVALITA 2011, pp. 86–97 (2013)
4. Basu, S., Yu, Y., Zimmermann, R.: Fuzzy clustering of lecture videos based on topic modeling. In: Proc. of the 14th Intl. Workshop on Content-Based Multimedia Indexing. pp. 1–6 (2016)
5. Cardoso-Cachopo, A.: Improving Methods for Single-label Text Categorization. PdD Thesis, Instituto Superior Tecnico, Universidade Tecnica de Lisboa (2007)
6. Chang, H., Kim, H., Li, S., Lee, J., Lee, D.: Comparative study on subject classification of academic videos using noisy transcripts. In: Proc. of the 4th IEEE Intl. Conf. on Semantic Computing. pp. 67–72 (2010)
7. Imran, A.S., Cheikh, F.A.: Blackboard content classification for lecture videos. In: 18th IEEE Intl. Conf. on Image Processing. pp. 2989–2992 (2011)
8. Jiang, H., Lu, Y., Xue, J.: Automatic soccer video event detection based on a deep neural network combined CNN and RNN. In: 28th IEEE Intl. Conf. on Tools with Artificial Intelligence. pp. 490–494 (2016)
9. Kapela, R., Swietlicka, A., Rybarczyk, A., Kolanowski, K., O'Connor, N.E.: Real-time event classification in field sport videos. Image Commun. 35(C), 35–45 (2015)
10. Lorenzo, B., Costantino, G., Cucchiara, R.: Neuralstory: an interactive multimedia system for video indexing and re-use. In: Proc. of the 15th Intl. Workshop on Content-Based Multimedia Indexing (2017)
11. Messina, A., Borgotallo, R., Dimino, G., Airola Gnota, D., Boch, L.: ANTS: A complete system for automatic news programme annotation based on multimodal analysis. In: 9th Intl. Workshop on Image Analysis for Multimedia Interactive Services. pp. 219–222 (2008)
12. Messina, A., Montagnuolo, M., Di Massa, R., Borgotallo, R.: Hyper media news: a fully automated platform for large scale analysis, production and distribution of multimodal news content. Multimedia Tools Appl. 63(2), 427–460 (2013)
13. Mocanu, B.C., Tapu, R., Zaharia, T.B.: Automatic segmentation of TV news into stories using visual and temporal information. In: Proc. of the 17th Intl. Conf. on Advanced Concepts for Intelligent Vision Systems. pp. 648–660 (2016)
14. Shih, H.: A survey on content-aware video analysis for sports. CoRR abs/1703.01170 (2017)
15. Xue, N., Xia, F., Chiou, F.d., Palmer, M.: The penn chinese treebank: Phrase structure annotation of a large corpus. Nat. Lang. Eng. 11(2), 207–238 (Jun 2005)
16. Yang, H., Oehlke, C., Meinel, C.: An Automated Analysis and Indexing Framework for Lecture Video Portal, pp. 285–294 (2012)
17. Yang, H., Siebert, M., Lhne, P., Sack, H., Meinel, C.: Lecture video indexing and analysis using video ocr technology. In: 7th Int. Conf. on Signal Image Technology and Internet Based Systems (SITIS 2011), Track Internet Based Computing and Systems (2011)
18. Zlitni, T., Bouaziz, B., Mahdi, W.: Automatic topics segmentation for tv news video using prior knowledge. Multimedia Tools Appl. 75(10), 5645–5672 (2016)