When you Doubt, Abstain: from Misclassification to Epoché in Automatic Text Categorisation

Angela Locoro
Computer Science Dep.
University of Genova, Italy
Locoro@disi.unige.it

Daniele Grignani
Modern and Contemporary History Dep.
University of Genova,Italy
Daniele.Grignani@gmail.com

Viviana Mascardi
Computer Science Dep.
University of Genova, Italy
Mascardi@disi.unige.it

***Abstract*—This paper describes how natural language processing and ontologies are exploited for automatic text categorisation. The approach introduced is part of the MANENT system, an infrastructure for integrating, structuring and searching Digital Libraries. The procedure of structural information extraction, and of the automatic classification of the records according to natural language understanding and the WordNet Domains taxonomy is discussed. A comparison between two versions of the classification algorithm is conducted and the improvements of the new approach are articulated. In particular, using semantic connections between words refines the classification results while reducing misclassification to non classification.**

***Keywords*-automatic text categorisation; natural language processing; wordnet domains; semantic digital libraries**

I. Introduction

Although the Web seems to represent the ultimate technology for transforming the process of knowledge proliferation and availability, it is more and more clear that the impressive amount of online resources often prevents the same expert users from a sensible choice and causes strong limitations in effective information acquisition.

Digital Libraries try to regulate the uncontrolled proliferation of information on the Web by preserving authoritative and well structured sources of knowledge. Research communities interoperate in a social network of interconnected digital libraries by sharing archival practices, methods and tools that rely on standard metadata formats and ontologies describing people, institutions and contents.

Many European projects[1] and recent research works [1], [2] witness the importance of a research field growing swiftly for the management of information commitment foreseen in the near future. The main emphasis of these researches is put on crucial aspects for the sustainability of new generation infrastructures such as the easiness of information discovery through natural language facets, the design of digital libraries more and more relying on trusted reference models and well grounded standard resource descriptions enriched with semantics, the provision of basic services such as contents personalisation, and the interoperability cornerstone.

In this scenario, the presence of efficient tools for storing and sharing all the needed theoretical and analytical knowledge becomes crucial, but without algorithms that ensure the *correct automatic categorisation* of documents belonging to digital libraries, all these tools and services might be useless, since users could not be able to filter and retrieve the information they are looking for.

This paper describes an algorithm for ontology-driven automatic metadata classification that greatly improves the correctness of the classification with respect to the original algorithm we designed, implemented and tested within the MANENT infrastructure [3]. The new algorithm is based on semantic correlation between words in metadata contents and relies on WordNet thesauri semantic relationships.

II. Background

In this section we briefly recall the methodologies and tools upon which our approach relies.

A. Digital Library Metadata access

The Metadata Harvester Service of the MANENT system is based on OAI-PMH [4] and retrieves records from worldwide repositories that contain Dublin Core [5] metadata. Data are downloaded in XML format as well as in plain text. An example of a downloaded record (we only cut the description field for space reasons), provided in plain text from the institutional repository of the University of Stirling, Scotland, is the following:

```
dc:title Republicanism and the American
Gothic
dc:subject Literature and culture United
States
dc:description This thesis explores the
republican tradition of the British
Enlightenment and the effect of its
translation and migration to the American
colonies.
```

where each metadata field is composed of the metadata label (e.g. *dc:title*) as well as the metadata content (e.g. "Republicanism and the American Gothic").

[1] See for example http://cordis.europa.eu/fp7/ict/telearn-digicult/, following the link to "DigiCult".

B. WordNet

In the WordNet Lexical Database [6] nouns, verbs, adjectives and adverbs are grouped into sets of cognitive synonyms (synsets). Most synsets are connected to other synsets according to semantic relations. In our approach we use the following ones: *hypernyms*: Y is a hypernym of X if every X is a (kind of) Y; *hyponyms*: Y is a hyponym of X if every Y is a (kind of) X; *coordinate terms*: Y is a coordinate term of X if X and Y share a hypernym; *holonym*: Y is a holonym of X if X is a part of Y; *meronym*: Y is a meronym of X if Y is a part of X.

C. WordNet Domains and the WordNet Domains Ontology

WordNet Domains[2] [7], [8] is a project developed by the Fondazione Bruno Kessler (FBK), Trento. The project has the aim of tagging WordNet synsets with domain labels. The domain labels are those of the Dewey Decimal Classification system[3], a standard largely adopted by library systems. The task of labelling synsets has been conducted on WordNet version 2.0. Mappings files from the oldest version to the newest exist and are freely available.

As part of our recent research [9], we took the WordNet Domains taxonomy[4] and codified it in OWL using Protégé. We called this new ontology WordNetDomains.owl. It consists of 160 domain labels divided as follows:

- 11 top level classes which represent the upper layer of domains classification. They are: *applied_science*, *doctrines*, *factotum*, *free_time*, *metrology*, *number*, *person*, *pure _science*, *quality*, *social_science*, *time_period*;
- 42 mid level classes that are used to tag synsets representing concepts used at an intermediate level of generality (e.g. *medicine*, *economy*, *sport*, and so on);
- 107 low level classes that are subclasses of one of the 42 mid level concepts or belong to a further level of specialisation and are also used to tag synsets, which are relative to concepts used in more specialised domains (e.g. *psychiatry*, *banking*, *athletics* and so on).

III. Metadata Classification according to WordNet and WordNet Domains

A detailed description of the main steps of our automatic classification algorithm can be found in [3]. Here we briefly recall the main steps of it and provide a detailed view of only the new part of the procedure in Section III-A, as an improving step in refining the automatic classification results which we will compare to the old ones and discuss in Section IV.

For each record and each metadata content:

1) We tokenise each word and tag it with the GATE[5] POS (part-of-speech) tagging. Once classified in this way we retain only noun words.

2) We lemmatise each noun in order to obtain its canonical word form using the WordNet dictionary.

3) We count the occurrences of each lemma in each metadata content and obtain the total number of times the lemma occurs into a metadata field.

4) We filter lemmas in the *dc:description* field by retaining only those words whose frequencies sum amounts to 50% of the total frequencies counts.

5) We apply to each set of words of each metadata field obtained as above the Semantic Connection Discovery approach (see the next Section) that would result in a recomputation of their frequencies based on their semantic relatedness with other words.

6) We assign WordNet domain labels to each lemma according to the mapping file

7) Besides the "direct tagging" with the WordNet domain associated with the given lemma (if any), we also tag lemmas with the super-domain labels of the domain labels just assigned by looking at the WordNet Domains ontology. What we obtain in the end is a set of domain labels associated with each lemma. Lemmas that were tagged with no domain label are eliminated.

8) We associate the frequencies of each word to the domain labels and we apply a weight to this number, according to the type of metadata field to which the domain labels belong to. In this case we weighted domain labels in *dc:subject* 1, domain labels in *dc:title* 0.5, and domain labels in *dc:description* 0.25.

9) We rank the domain labels according to the resulting score. Each relevant top domain together with its sub-domains, if present, will appear in the ranking. The *Factotum* domain is filtered out as having no significance for the classification task we pursue.

A. The semantic connection discovery approach

For each document to be classified based on its metadata, as a refining step of the metadata contents preprocessing stage, and before the tagging of words with WordNet Domains labels, we take each word w from each list of words obtained for that document until phase 4 of the above procedure (one list of words for each metadata field) and compare it with any other word in the same list, and with any other one in the other lists. The comparison is conducted according to the following criteria:

- check if w is equal to any other word contained in any other metadata word lists;

- retrieve the synsets from WordNet, and for each synset, the hypernyms, hyponyms, meronyms and holonyms synsets of each word w in each of the metadata word lists; call

[2]http://wndomains.fbk.eu/index.html.

[3]http://www.oclc.org/dewey/versions/default.htm.

[4]Available at the official page of the project: http://wndomains.fbk.eu/hierarchy.html. Last accessed on 30 March 2011.

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

this WordNet set of synsets and of its semantically related synsets $WordNetConnected_w$

- compare each couple of words w_1 and w_2 by comparing the $WordNetConnected_{w_1}$ and $WordNetConnected_{w_2}$ lists. The result of this comparison would yield to four different results:

- w_1 and w_2 are *synonyms*
- w_1 is one of the hypernyms, hyponyms, meronyms or holonyms of w_2 or vice versa (we say that w_1 and w_2 are *directly related*)
- w_1 and w_2 have hypernyms, hyponyms, meronyms or holonyms in common (we say that w_1 and w_2 are *indirectly related*)
- w_1 and w_2 neither are equal nor have semantically connected synset in common (we say that w_1 and w_2 are *not related*)

- assign a score to each of the different results, depending on the two words relatedness (e.g. two equal words would receive a higher score than two words connected through a common hyponym)

- use this score for tuning the frequencies of the words calculated until step 4 of the above procedure by multiplying scores with frequencies. In this way, the frequency of words that are more semantically connected becomes higher with respect to less connected words frequency, while words that are not connected are discarded. For testing our approach we have used heuristic scores. For equal words and synonyms we assign a score of 3, for directly related words we assign a score of 2, for indirectly related words the score assigned is 1, and for non related words the score assigned is 0.

The rationale behind such approach is that an inter-metadata as well as an intra-metadata comparison of words characterising the contents allows the most meaningful terms with respect to the topic treated to emerge, as one domain of discourse is described by humans using a set of words that are semantically related to the argument being described with respect to other ones that are just part of the sentence for grammatical and style reasons. If we apply a technique to discover such highly semantic connected words and distill them, we have more chances to obtain the most significant keywords of the content itself, which may contribute to the correct automatic classification task.

IV. Experiments and Results

Our experiments have been conducted on a dataset of 10 repositories, chosen among the 1.342 all over the world repositories that expose OAI-PMH services. For each repository, we selected 10 records in the temporal range January 2008 - October 2010 and we asked domain experts to manually verify the correctness and completeness of our automatic classification procedure.

The selection of the 10 repositories was based on the completeness of the Dublin Core metadata available for each record (i.e., most records of the 10 repositories have at least one *dc:title*, one *dc:subject* and one *dc:description* field). The 100 records that were classified and manually inspected (our benchmark) are all in English.

Table I
Manual evaluation over the benchmark carried out by domain experts

	No_SC			With_SC				
rep.	♯ dom.	avg	∃	♯ dom.	avg	∀	impr	NC
r1	61	6.1	100%	40	4.0	30%	50%	1
r2	56	5.6	90%	40	4.0	30%	40%	
r3	42	4.2	80%	30	3.0	40%	40%	2
r4	66	6.6	100%	26	2.6	40%	40%	3
r5	42	4.2	90%	25	2.5	30%	60%	2
r6	40	4.0	60%	32	3.2	30%	30%	
r7	55	5.5	70%	39	3.9	40%	60%	1
r8	45	4.5	80%	35	3.5	50%	70%	
r9	48	4.8	90%	39	3.9	10%	30%	2
r10	63	6.3	90%	36	3.6	50%	50%	

Once the automatic classification procedure has been run we ranked the results obtained for each record according to the following mechanism:

- for the top level WordNet Domains with first two higher scores, search their direct sub-domains at the same ranking level down to the third ranking score;
- do the same for their leaf domains if they exist and if the scoring function has ranked them among the first three ranking scores.

In our previous experiments the average of tags for each record was about 5 (with 84% of records tagged with at least 2 top level domain labels). Top level domain labels are the most important ones for the correct classification, as they represent the path to more specific domains. Having a document tagged with 2 or more top level domains does not help the classification process since such an abundance of top level tags may often carry the same information as no tags at all. This noise (that is wrong labels) should be hence reduced, and documents should be tagged with fewer (and eventually all correct!) top level domains.

In table I we report the results of experiments conducted on both approaches (the one we developed as part of our past activity, without the semantic connection discovery, named "No_SC" in the table, and the new one with the semantic connection discovery, named "With_SC") for the benchmark records of the 10 repositories (r1, r2, and so on). Columns **"♯ dom."** and **"M"** represent the total number of domain labels extracted from our procedure and the average number of domain labels for record respectively, column ∃ in "No_SC" reports the percentage of records that were tagged with *at least one correct domain* (there might be a lot of noise however, if all the other domains were wrong), whereas column ∀ in "With_SC" reports the percentage of records whose top level domains *were all correct*. We do not include a ∀ column for the "NO_SC" case as well, since records that were tagged with only correct top level domains with that algorithm were a negligible percentage.

Column **"impr"** reports the overall improvement of "With_SC" with respect to "No_SC", in terms of *noise reduction*, and is the sum of the $\forall$ values plus the percentage of those records where the elimination of wrong labels was only partial. Finally, column **"NC"** shows the number of non classified records. Non classification only occurs when semantic connection discovery is applied, and we interpret it as a more sophisticated discrimination of data able to avoid the wrong choice in presence of insufficient data or not self-explaining data, a situation in which also a human being would raise her doubts during the manual evaluation.

Despite the encouraging result we have to observe that, in 7% of the cases, wrong classification occurred even when "With_SC" has been applied, and on records that, with "No_SC", had their correct top level domain assigned (together with other noisy top level domains). These bad results are not compromising the improvements obtained (the percentage of records misclassified is comparable with that of the previous approach, which is 5%). On the other hand we achieved an average improvement of 47% on the whole benchmark for noise elimination over correctly classified records. The number of records with only one correct top level label has raised to 35%, against a 16% with the "No_SC" algorithm.

Among the 11 non classified records, 5 contained only one metadata field out of three (so data were incomplete) with 3 of them containing only 1, 2 and 3 words respectively (so data were insufficient), 1 having a very brief description that results to be difficult to tag correctly by humans, whereas the last one has been tagged with 5 top level labels by "No_SC", and hence the problem is probably its cross-domain characterisation. 3 out of the 11 records were wrongly classified with "No_SC". In 1 record the right domain (namely *computer_science*) was difficult to be detected automatically due to lacks of competency words inside WordNet thesaurus for that specific domain. Finally, 2 records have not been classified despite the quite complete set of metadata and the number of words which were sufficient to try a classification. One reason is the experimented difficulty of WordNet in tagging records of the *agriculture* domain and records that overlap between *physics* and *chemistry* domains respectively.

Just to make an example, one of the non classified records consists of the following metadata:

```
dcsubject Birthmothers -- Correspondence
```

The domain experts were not able to classify it, hence the non classification result of our algorithm can be considered as a correct result.

V. Conclusion

In this paper we have presented an algorithm for automatic text classification according to WordNet Domains that was tested on MANENT system, an infrastructure for Digital Libraries able to download metadata records that are stored in online Digital Libraries all over the world. The classification algorithm has been improved with an approach called "Semantic Connection Discovery" and new experiments and results have been presented and discussed.

By considering non classified records as a better result with respect to misclassification, we intend to extend our manual evaluation in order to widely analyse under which conditions (lack of which information, metadata, malformed descriptions) our automatic system does not classify them. In particular, an investigation on the overlapping of such records with those that even a human being is not able to express a judgment over is going to be carried on. This can also be used to highlight to users that are in charge of filling the metadata content that the information they are giving is not at all complete, and this can be an aid to a better metadata fulfillment for describing online resources.

We are confident that this new direction is worth a deeper analysis and further improvements; for example, the exploitation of techniques that mix the two approaches, in order to keep the completeness while not compromising the correctness, will be further investigated.

References

[1] A. Baruzzo, P. Casoto, P. Challapalli, A. Dattolo, N. Pudota, and C. Tasso, "Toward semantic digital libraries: Exploiting Web2.0 and semantic services in cultural heritage," *Journal of Digital Information*, vol. 6, no. 10, pp. 1047–1053, 2009.

[2] S. Kruk and B. McDaniel, *Semantic Digital Libraries*. Springer, 2009.

[3] A. Locoro, D. Grignani, and V. Mascardi, *MANENT: An Infrastructure for Integrating, Structuring and Searching Digital Libraries*, ser. Studies in Computational Intelligence. Springer, 2011, ch. to be published.

[4] "Oai-pmh: Open archives initiative protocol for metadata harvesting." [Online]. Available: http://www.openarchives.org/OAI/openarchivesprotocol.html

[5] "Dublin core metadata element set." [Online]. Available: http://www.dublincore.org/documents/dces/

[6] G. Miller, "Wordnet: A lexical database for english," *Communications of the ACM*, vol. 11, no. 38, pp. 39–41, 1995.

[7] B. Magnini and G. Cavagliá, "Integrating Subject Field Codes into WordNet," in *Proc. of the 2nd International Conference on Language Resources and Evaluation, (LREC-2000)*, M. Gavrilidou, G. Crayannis, S. Markantonatu, S. Piperidis, and G. Stainhaouer, Eds., 2000, pp. 1413–1418.

[8] L. Bentivogli, P. Forner, B. Magnini, and E. Pianta, "Revising WordNet domains hierarchy: Semantics, coverage, and balancing," in *Proc. of the 21st International Conference on Computational Linguistics, (COLING 2004)*, 2004, pp. 101–108.

[9] A. Locoro, "Tagging ontologies with fuzzywordnet domains," in *Proc. 9th International Workshop on Fuzzy Logic and Applications (WILF'2011)*, ser. LNCS. Trani, Bari: Springer, Aug. 2011, p. to appear.