MOoD-TC, a Multilingual Ontology Driven Text Classifier

Angelo Ferrando, Silvio Beux
and **Viviana Mascardi**
DIBRIS, University of Genova,
Via Dodecaneso 35, 16146,
Genova, ITALY
angelo.ferrando@dibris.unige.it
silviobeux@gmail.com
viviana.mascardi@unige.it

Paolo Rosso
PRHLT, Universitat Politècnica de València,
Campus de Vera, 46022,
València, SPAIN
prosso@dsic.upv.es

Abstract

In this paper we present a Multilingual Ontology-Driven framework for Text Classification (MOoD-TC). This framework is highly modular and can be customised to create applications based on Multilingual Natural Language Processing for classifying domain-dependent contents.

1 Introduction

The large amount of digital data made available in the last years from a wide variety of sources raises the need for automatic methods to extract meaningful information from them. The extracted information is precious for many purposes, and especially for commercial ones. Jackson and Moulinier (Jackson and Moulinier, 2002) observe that *"there is no question concerning the commercial value of being able to classify documents automatically by content. There are myriad potential applications of such a capability for corporate Intranets, government departments, and Internet publishers"*.

The problem of classifying multilingual pieces of text was addressed since the end of the last millennium (Oard and Dorr, 1996) but it is still a significant problem because each language has its own peculiar features, making the automatic management of multilingualism an open issue.

The use of ontologies to classify multilingual texts (de Melo and Siersdorfer, 2007) is a good alternative to standard machine learning approaches in all those situations where a training set of documents is not available or it is too small to properly train the classifier. Ontology-driven text classification does not depend on the existence of a training set, as it relies solely on the entities, their relationships, and the taxonomy of categories represented in an ontology, that becomes the driver of the classification. Another advantage of ontology-driven classification is that ontology concepts are organised into hierarchies and this makes possible to identify the category (or the categories) that best classify the document's content, by traversing the hierarchical structure.

In this paper we present MOoD-TC (*M*ultilingual *O*ntology *D*riven Text Classifier (Beux, 2015; Leotta et al., 2015)), a highly modular system which has been conceived, designed and implemented to be customised by the system developer for obtaining different domain-dependent behaviours, always centered around the multilingual text classification process.

The paper is organised as follows: Section 2 describes MOoD-TC, Section 3 shows some results obtained using MOoD-TC, Section 4 analyses the state of the art, and Section 5 concludes.

2 MOoD-TC

The *M*ultilingual *O*ntology *D*riven Text Classifier (MOoD-TC) has been developed as part of Silvio Beux' Masters Thesis (Beux, 2015), starting from (Leotta et al., 2015). Its aim is to classify multilingual textual documents according to classes described in a domain ontology. MOoD-TC consists of the Text Classifier (TC) and the Application Domain Module (ADM).

It provides a set of core modules offering functionalities which are common to any text classification problem (text pre-processing, tagging, classification) plus a customisable structure for those modules which can be implemented by the developer in order to offer application-specific functionalities. It returns a classification of the text w.r.t. the ontology taken as input. The classification performed by TC which is implemented in Java and exploits the Language Detector Library[1], Babel-

[1] https://code.google.com/p/language-detection/

Net (Navigli and Ponzetto, 2012), and TreeTagger[2].

Since we are interesting more in the use of BabelNet to solve the multilingual aspect, we report in Figure 1 just an overview of the TC behaviour.

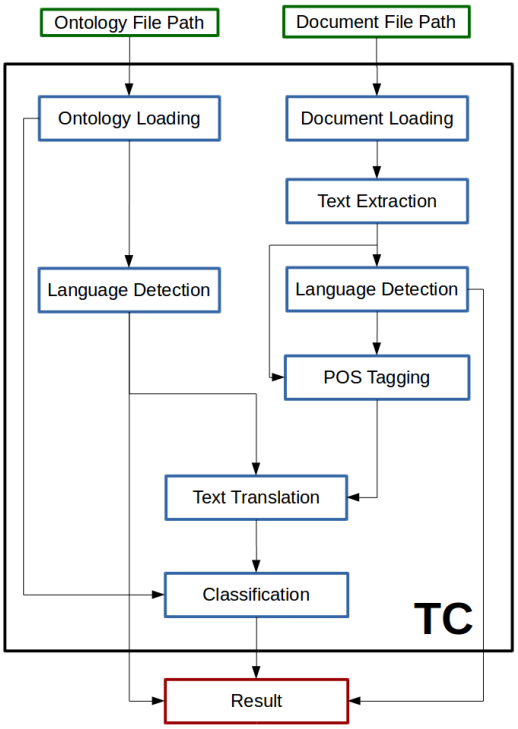

Figure 1: High level design of TC.

The Language Detector Library detects, with a precision greater than 99%, 53 languages making use of Naive Bayesian filters. It is devoted to recognise the language L_o of the ontology o and the language L_d of the textual document d. This step is necessary because in our process pipeline we need to know the language of the inputs (o and d) to perform the multilingual classification. The TreeTagger tool performs the tagging of d in order to obtain, for each word $w \in d$ different from a stop word, its lemma (the canonical form of the word) and its part of speech (POS). This information, the tuple *(lemma, POS)*, is used by BabelNet to perform the translation of w into the ontology language in this way: for each tuple, all the *synsets* in the text language containing the lemma and the *POS* of the original word w contained in the tuple are retrieved from BabelNet. To do this, we used the BabelNet function *getSynsets*. Once we obtain the synsets, we have to take all the senses associated with each synset in the language of our ontology. Given L_o the target language used in the ontology, all the words associated with each synset $s \in S$ in the language L_o are retrieved by means of the BabelNet function *getSenses*. At this point, for each translated sense, we check if it is present into the ontology model M_o previously stored. Then the ontology is scanned concept by concept and, if a match is found between the sense and the concept, the search stopped (it is also taken into account the number of occurrences of each concept found in the text).

The *ClassifierObject* is the object that stores a correctly classified word (and additional information) of the document d with respect to o. TC returns a list of such objects.

2.1 ADM

ADM specialises the text classifier task by implementing functionalities for pre- and post- processing a multilingual textual document. If an ADM is used, the entire system specialises its behaviour in the domain represented by that particular ADM. In our system TC can work alone, but an ADM is meant to work in close connection with the core system.

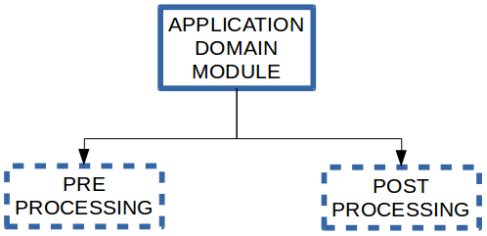

Figure 2: ADM.

The core modules are implemented to work for the European languages (which share some common features like, for example, the relationship between noun and adjective), but they could be extended to cope with the peculiar features of other languages; in fact, thanks to the modularity of the system, it is possible to integrate different algorithms created specifically to handle that peculiarities, without modifying the entire system. The ADM processes the TC input and output in order to obtain a new domain oriented tool. An ADM is composed by two subcomponents: pre-processing and post-processing. The pre-processing component takes as input a digital object and returns a new processed text

[2] http://code.google.com/

while the post-processing component takes as input the TC output and returns a domain dependent result. Figure 3 shows the entire pipeline of the integration process between the TC and the ADM.

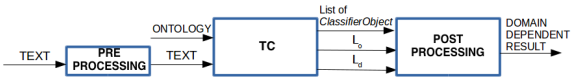

Figure 3: Integration pipeline of TC and ADM.

3 Experiments

We used MOoD-TC in two different scenarios: Sentiment Analysis and Symptoms Identification.

In both cases we reused the core of our framework, the TC, to classify with respect to a domain ontology. In the first, the ontology concerned sentiments which could be expressed by users with regard to an hotel. In the second, the ontology contained concepts related to disease symptoms.

The main difference between the two implementations concerns the pre- and post- processing phases and the consequently customisation of the ADM. For sentiment analysis, after the TC returns the classified features extracted from the text, it evaluates the polarity of the review on the base of polarity of each single feature found. Instead, for symptoms identification, with the selected features it can conclude a set of possible diseases on the base of all symptoms identified.

Thanks to the modularity of our approach, both the implementations were implemented in a very easy way simply changing the ADM. This is due to the TC robustness and generality obtained through BabelNet, which allows managing a wide range of languages indifferently.

We would mention the results obtained in the symptoms identification in order to show the powerfulness of our text classifier helped by BabelNet.

The experiments have been carried out on 32 sentences for each of the 5 languages (English, French, German, Italian, Spanish), for a total of 160 sentences. Each sentence describes symptoms related to one of the following sixteen disease: tinnitus, food allergy, cervical, dehydration, hyperthyroidism, flu, appendicitis, food poisoning, labyrinthitis, narcolessia, pneumonia, diabetes type 1, hyperglycemia, hypoglycemia, bronchitis, jet lag (two sentences for each disease). These diseases were chosen totally randomly in order to be closer at a real case where the diseases could be of each type. To cover the widest range of cases we also tried to consider the diseases with the most varied symptoms. The description of symptoms associated with each disease have been retrieved from (Griffith, 1992) and each sentence contains 2 up to 9 symptom words.

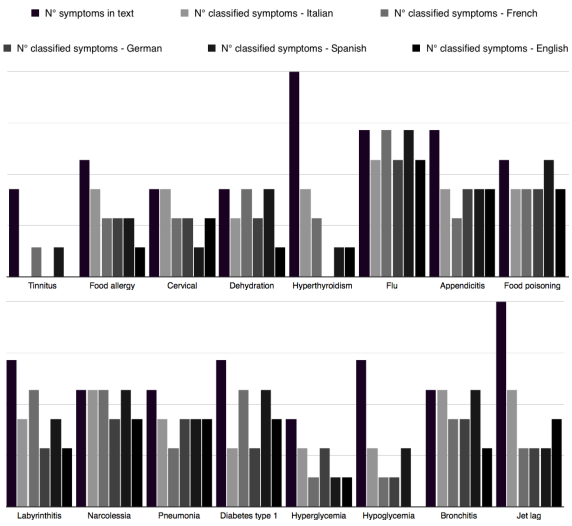

Figure 4: For each disease, the corresponding number of symptoms classified for each tested language.

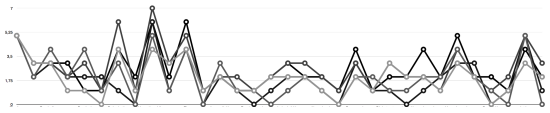

Figure 5: Trend of errors for disease in the five languages (False Negatives).

Figure 4 shows the results of our experiments in counting false negatives in the sentences representative of each disease. For each disease, the first column (in black) measures the average number of symptoms that should have been identified; the next five columns show the average number of correctly identified symptoms in Italian, French, German, Spanish and English sentences respectively. Figure 5 shows the number of false negatives (y axis) for disease (x axis, labels are omitted as they would not be readable) and for each language: each line is associated with one language. Figure 4 shows that the results greatly vary with the disease.

4 State of the art

Referring to the work of (Bentaallah and Malki, 2012), in order to overcome the disadvantage of using machine translation systems, many re-

searches have been working on using linguistic resources such as bilingual dictionaries and comparable corpora to induce correspondences between two languages. A good result obtained by (Bentaallah and Malki, 2012) is the answer to the question: *"The use of WordNets*[3] *in Text Categorisation guarantee good results better than those obtained by the Princeton WordNet?"*. The authors show through experiments that the answer to this question is *No*, or rather, the use of WordNets does not guarantee good results rather than those obtained by the Princeton WordNet (using a machine translation). This result leads us to believe that also in our case, where we have a single monolingual ontology, the results should not be so different with respect to those obtained through the use of a multilingual ontology, where there are concepts replicated for each language.

These last years, researches showed that using ontologies in monolingual text categorisation is a promising track. In (Guyot et al., 2005) authors propose a new approach that consists in using a multilingual ontology for Information Retrieval, without using any translation. Our work has similar properties, total absence of translation, but does not solve the multilingual problem through the use of a multilingual ontology but using a easier monolingual ontology supported by a multilingual semantic network as BabelNet, which solves the semantic correlation among words expressed in different languages.

Another interesting work using a really different approach in the field of intelligent methods are proposed in (Chau et al., 2005), where the concept-based hierarchical Multilingual Text Categorisation is enabled through the use of neural networks, and in (Gliozzo and Strapparava, 2006), where the authors propose a new approach to solve the Multilingual Text Categorisation problem based on acquiring Multilingual Domain Models from comparable corpora to define a generalised similarity function (i.e. a kernel function) among documents in different languages, which is used inside a Support Vector Machines classification framework.

5 Conclusions and Future Work

Our framework does not face many well known open problems in multilingual text classification and information extraction such as negation [22] and named entities, but rather it provides a flexible and modular approach ready for integrating, with limited effort, the results and algorithms addressing the above problems coming from the research community.

Since BabelNet is a very powerful and easy to use multilingual semantic network, our future works on this front will be targeted to:

- improve the POS processing in order to have more detailed and useful information which allow optimizing the search within BabelNet;
- manage the negation, the irony and other important POS aspects which have to be resolved in order to query BabelNet correctly;
- test our classifier on different scenarios in order to validate and check its potential, we already tested it in two different scenarios to solve a Sentiment Analysis and a Symptoms Extraction problem.

[3] WordNets are language customised version of the Princeton WordNet, which is in english.

References

Mohamed Amine Bentaallah and Mimoun Malki. 2012. The use of wordnets for multilingual text categorization: A comparative study. In *Proceedings of the 4th International conference on Web and Information Technologies, ICWIT 2012, Sidi Bel Abbes, Algeria, April 29-30, 2012*, pages 121–128.

Silvio Beux. 2015. MOoD-TC: A general purpose multilingual ontology driven text classifier. Master's Degree Thesis in Computer Science, University of Genova, Italy.

Rowena Chau, Chung-Hsing Yeh, and Kate A. Smith. 2005. A neural network model for hierarchical multilingual text categorization. In *Advances in Neural Networks - ISNN 2005, Second International Symposium on Neural Networks, Chongqing, China, May 30 - June 1, 2005, Proceedings, Part II*, pages 238–245.

Gerard de Melo and Stefan Siersdorfer. 2007. Multilingual text classification using ontologies. In Giambattista Amati, Claudio Carpineto, and Giovanni Romano, editors, *Advances in Information Retrieval, 29th European Conference on IR Research, ECIR 2007, Rome, Italy, April 2-5, 2007, Proceedings*, volume 4425 of *Lecture Notes in Computer Science*, pages 541–548. Springer.

Alfio Massimiliano Gliozzo and Carlo Strapparava. 2006. Exploiting comparable corpora and bilingual dictionaries for cross-language text categorization. In *ACL 2006, 21st International Conference on*

Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference, Sydney, Australia, 17-21 July 2006.

H. Winter Griffith. 1992. *Complete guide to symptoms, illness & surgery for people over 50 / H. Winter Griffith ; surgical illustrations by Mark Pederson*. Body Press/Perigee New York, NY.

Jacques Guyot, Saïd Radhouani, and Gilles Falquet. 2005. Ontology-based multilingual information retrieval. In *Working Notes for CLEF 2005 Workshop co-located with the 9th European Conference on Digital Libraries (ECDL 2005), Wien, Austria, September 21-22, 2005.*

Peter Jackson and Isabelle Moulinier. 2002. *Natural Language Processing for Online Applications: Text Retrieval, Extraction & Categorization*. John Benjamins.

Maurizio Leotta, Silvio Beux, Viviana Mascardi, and Daniela Briola. 2015. My MOoD, a multimedia and multilingual ontology driven MAS: design and first experiments in the sentiment analysis domain. In *Proceedings of the 2nd International Workshop on Emotion and Sentiment in Social and Expressive Media: Opportunities and Challenges for Emotion-aware Multiagent Systems*, pages 51–66.

Roberto Navigli and Simone Paolo Ponzetto. 2012. Babelnet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artif. Intell.*, 193:217–250.

Douglas W. Oard and Bonnie J. Dorr. 1996. A survey of multilingual text retrieval. Technical report, College Park, MD, USA.