International Journal Of Business Management
Available at www.ijbm.co.in
ISSN NO. 2349-3402
VOL. 2(2),2015

CROSS LANGUAGE TEXT CLASSIFICATION

Ajay Gupta[1]

Abstract

The automated categorization (or classification) of texts into predefined categories has witnessed a booming interest in the last ten years, due to the increased availability of documents in digital form and the ensuing need to organize them.

Due to the globalization on the Web, many companies and institutions need to efficiently organize and search repositories containing multilingual documents. The management of these heterogeneous text collections increases the costs significantly because experts of different languages are required to organize these collections. Cross-Language Text Categorization can provide techniques to extend existing automatic classification systems in one language to new languages without requiring additional intervention of human experts.

In the proposed approach, we assume that a predefined category set and a collection of labelled training data is available for a given language.

1. SAHEED UDHAM SINGH GROUP OF COLLEGE, TANGORI, MOHALI

International Journal Of Business Management
Available at www.ijbm.co.in
ISSN NO. 2349-3402
VOL. 2(2),2015

INTRODUCTION

The most popular and successful algorithms for text classification are based on machine learning techniques. Learning systems have the advantage of flexibility since the only required human effort is to provide a consistent set of labelled examples. However, labelling a dataset can be costly because different expertises may be required. In fact, in recent years also because of the growth in the popularity of the Web, many companies and organizations were required to manage documents in different languages. Multi-Language Text Classification became an important task. This task can be approached as a set of Mono-Language problems, but a different labelled dataset would be needed to train a classifier for each language. This labelling process can be quite costly since it needs an expert for each different language.

Text categorization the activity of labelling natural language texts with thematic categories from a predefined set, is one such task. TC dates back to the early ’60s, but only in the early ’90s it became a major subfield of the information systems discipline, thanks to increased applicative interest and to the availability of more powerful hardware. TC is now being applied in many contexts, ranging from document indexing based on a controlled vocabulary, to document filtering, automated metadata generation, word sense disambiguation, population of hierarchical catalogues of Web resources, and in general any application requiring document organization or selective and adaptive document dispatching.

International Journal Of Business Management
Available at www.ijbm.co.in
ISSN NO. 2349-3402
VOL. 2(2),2015

Cross-Language Text Categorization (CLTC) is a new area in text categorization. The CLTC task can be stated as follows: suppose we have a good classifier for a set of categories in a language M1 and a large amount of unlabeled data in a different language M2; how can we categorize this corpus according to the same categories defined for language M1 without having to manually label any data in M2? When using the machine learning paradigm, this problem can be reformulated as: how can we train a text classifier for language M2 using the examples labelled for language M1? An algorithm that is able to effectively perform this task would reduce the costs of building multi-language classification systems, since the human effort would be reduced to provide a training set in just one language. Until the late ’80s the most popular approach to TC, at least in the “operational” (i.e. real-world applications) community, was a knowledge engineering (KE) one, consisting in manually defining a set of rules encoding expert knowledge on how to classify documents under the given categories. In the ’90s this approach has increasingly lost popularity in favour of the machine learning (ML) paradigm, according to which a general inductive process automatically builds an automatic text classifier by learning, from a set of pre classified documents, the characteristics of the categories of interest. The advantages of this approach are accuracy comparable to that achieved by human experts, and a considerable savings in terms of expert manpower, since no intervention from either knowledge engineers or domain experts is needed for the construction of the classifier or for its porting to a different set of categories. It is the ML approach to TC that this paper concentrates on.

International Journal Of Business Management
Available at www.ijbm.co.in
ISSN NO. 2349-3402
VOL. 2(2),2015

Cross-Language Text Classification

For reason of simplicity, we reduce the multi-lingual case with k languages to k − 1 bi-lingual problems selecting one language as the principal one; thus studying the bi-lingual case is not restrictive with respect to the multilingual problem. Before describing some aspects of the

Cross-Lingual Text Categorization task, we introduce some notations used in the following sections. We denote the two languages with L1 and L2 and with L2!1 the language L1generated by the translation from L2. Moreover, we denote with TR1 and TS1 the training set and the test set in language L1, and with TR2!1 and TS2!1 the training set TR2 and the test set TS2 translated into the language L1.

TEXT CATEGORIZATION

Text categorization is the task of assigning a Boolean value to each pair hdj , cii ∈ D×C, where D is a domain of documents and C = {c1, . . . , c|C|} is a set of predefined categories. A value of T assigned to hdj , cii indicates a decision to file dj under ci, while a value of F indicates a decision not to file dj under ci. More formally, the task is to approximate the unknown target function ˘_ : D ×C → {T, F} (that describes how documents ought to be classified) by means of a function _ : D × C → {T, F} called the classifier (aka rule, or hypothesis, or model) such that ˘_ and _ "coincide as much as possible". we will assume that: —The categories are just symbolic labels, and no additional knowledge (of a procedural or declarative nature) of their meaning is available. No exogenous knowledge (i.e. data provided for classification purposes by an external source)

International Journal Of Business Management
Available at www.ijbm.co.in
ISSN NO. 2349-3402
VOL. 2(2),2015

is available; therefore, classification must be accomplished on the basis of endogenous knowledge only (i.e. knowledge extracted from the documents). In particular, this means that metadata such as e.g. publication date, document type, publication source, etc. is not assumed to be available.

APPLICATIONS OF TEXT CATEGORIZATION

It has been used for a number of different applications, of which we here briefly review the most important ones. Note that the borders between the different classes of applications listed here are fuzzy and somehow artificial, and some of these may be considered special cases of others. Applications are speech categorization by means of a combination of speech recognition and multimedia document categorization through the analysis of textual captions , author identification for literary texts of unknown or disputed authorship, language identification for texts of unknown, automated identification of text genre, and automated essay grading

THE MACHINE LEARNING APPROACH TO TEXT CATEGORIZATION

In the ’80s the most popular approach for the creation of automatic document classifiers consisted in manually building, by means of knowledge engineering (KE) techniques, an expert system capable of taking TC decisions. Such an expert system would typically consist of a set of manually defined logical rules, one per category, of type if DNF formula then hcategoryi A DNF (“disjunctive normal form”) formula is a disjunction of conjunctive clauses; the document is classified under category if it satisfies the formula, i.e. if it satisfies at least one of the clauses.

International Journal Of Business Management
Available at www.ijbm.co.in
ISSN NO. 2349-3402
VOL. 2(2),2015

if ((wheat & farm) or (wheat & commodity) or (bushels & export) or (wheat & tonnes) or

(Wheat & winter & ¬ soft)) Then Wheat else ¬ Wheat. Rule-based classifier for the Wheat category; keywords are indicated in italic, categories are indicated in Small Caps. The Construe system , built by Carnegie Group for the Reuters news agency. The drawback of this approach is the knowledge acquisition bottleneck well-known from the expert systems literature. That is, the rules must be manually defined by a knowledge engineer with the aid of a domain expert if the set of categories is updated, then these two professionals must intervene again, and if the classifiers ported to a completely different domain (i.e. set of categories) a different domain expert needs to intervene and the work has to be repeated from scratch. On the other hand, it was originally suggested that this approach can give very good effectiveness results:. In ML terminology, the classification problem is an activity of supervised learning, since the learning process is "supervised" by the knowledge of the categories and of the training instances that belong to them. The advantages of the ML approach over the KE approach are evident. The engineering effort goes towards the construction not of a classifier, but of an automatic builder of classifiers . This means that if a learner is available off-the-shelf, all that is needed is the inductive, automatic construction of a classifier from a set of manually classified documents. The same happens if a classifier already exists and the

International Journal Of Business Management
Available at www.ijbm.co.in
ISSN NO. 2349-3402
VOL. 2(2),2015

original set of categories is updated, or if the classifier is ported to a completely different domain.

Training set, test set, and validation set

The ML approach relies on the availability of an initial corpus = {d1, . . . , d||} D of documents pre classified under C = {c1, . . . , c|C|}. That is, the values of the total function ˇ_ : D × C → {T, F} are known for every pair hdj , cii ∈ × C. A document dj is a positive example of ci if ˇ_(dj , ci) = T , a negative example of ci if ˇ_(dj , ci) = F.In research settings once a classifier _ has been built it is desirable to evaluate its effectiveness. In this case, prior to classifier construction the initial corpus is split in two sets, not necessarily of equal size: —a training(-and-validation) set T V = {d1, . . . , d|TV |}. The classifier _ for categories C = {c1, . . . , c|C|} is inductively built by observing the characteristics of these documents; —a test set T e = {d|TV |+1, . . . , d||}, used for testing the effectiveness of the classifiers. Each dj ∈ T e is fed to the classifier, and the classifier decisions _(dj , ci)are compared with the expert decisions ˇ_(dj , ci). A measure of classification effectiveness is based on how often the _(dj , ci) values match the ˇ_(dj , ci) values. The documents in T e cannot participate in any way in the inductive construction of the classifiers; if this condition were not satisfied the experimental results obtained would likely be unrealistically good, and the evaluation would thus have no scientific character. In an operational setting, after evaluation has been performed one would typically re-train the classifier on the entire initial corpus, in order to boost effectiveness. In this case the results of the previous evaluation would be a pessimistic estimate of the real performance, since the final classifier has been

International Journal Of Business Management
Available at www.ijbm.co.in
ISSN NO. 2349-3402
VOL. 2(2),2015

trained on more data than the classifier evaluated. This is called the train-and-test approach

Information retrieval techniques and text categorization

Text categorization heavily relies on the basic machinery of information retrieval (IR). The reason is that TC is a content-based document management task, and as such it shares many characteristics with other IR tasks such as text search.

IR techniques are used in three phases of the text classifier life cycle:

(1) IR-style indexing is always performed on the documents of the initial corpus and on those to be classified during the operational phase.

(2) IR-style techniques are often used in the inductive construction of the classifiers

(3) IR-style evaluation of the effectiveness of the classifiers is performed.

Conclusions

We presented a new technique to categorize text documents in a cross-language environment. It is motivated by the availability of documents written in different languages and also by the fact that companies need to build categorization systems for these multi-lingual documents.

Manually labelling a large number of documents in each language is very labour intensive and time consuming