Cross language Text Categorization by acquiring Multilingual Domain Models from Comparable Corpora

Alfio Gliozzo and **Carlo Strapparava**
ITC-Irst
via Sommarive, I-38050, Trento, ITALY
{gliozzo,strappa}@itc.it

Abstract

In a multilingual scenario, the classical monolingual text categorization problem can be reformulated as a *cross language TC* task, in which we have to cope with two or more languages (e.g. *English* and *Italian*). In this setting, the system is trained using labeled examples in a source language (e.g. *English*), and it classifies documents in a different target language (e.g. *Italian*).

In this paper we propose a novel approach to solve the cross language text categorization problem based on acquiring Multilingual Domain Models from comparable corpora in a totally unsupervised way and without using any external knowledge source (e.g. bilingual dictionaries). These Multilingual Domain Models are exploited to define a generalized similarity function (i.e. a kernel function) among documents in different languages, which is used inside a Support Vector Machines classification framework. The results show that our approach is a feasible and cheap solution that largely outperforms a baseline.

1 Introduction

Text categorization (TC) is the task of assigning category labels to documents. Categories are usually defined according to a variety of topics (e.g. SPORT, POLITICS, etc.) and, even if a large amount of hand tagged texts is required, the state-of-the-art supervised learning techniques represent a viable and well-performing solution for monolingual categorization problems.

On the other hand in the worldwide scenario of the web age, multilinguality is a crucial issue to deal with and to investigate, leading us to reformulate most of the classical NLP problems. In particular, monolingual Text Categorization can be reformulated as a *cross language TC* task, in which we have to cope with two or more languages (e.g. *English* and *Italian*). In this setting, the system is trained using labeled examples in a source language (e.g. *English*), and it classifies documents in a different target language (e.g. *Italian*).

In this paper we propose a novel approach to solve the cross language text categorization problem based on acquiring Multilingual Domain Models (MDM) from comparable corpora in an unsupervised way. A MDM is a set of clusters formed by terms in different languages. While in the monolingual settings semantic domains are clusters of related terms that co-occur in texts regarding similar topics (Gliozzo et al., 2004), in the multilingual settings such clusters are composed by terms in different languages expressing concepts in the same semantic field. Thus, the basic relation modeled by a MDM is the domain similarity among terms in different languages. Our claim is that such a relation is sufficient to capture relevant aspects of topic similarity that can be profitably used for TC purposes.

The paper is organized as follows. After a brief discussion about comparable corpora, we introduce

Proceedings of the ACL Workshop on Building and Using Parallel Texts, pages 9–16,
Ann Arbor, June 2005.

a multilingual Vector Space Model, in which documents in different languages can be represented and then compared. In Section 4 we define the MDMs and we present a totally unsupervised technique to acquire them from comparable corpora. This methodology does not require any external knowledge source (e.g. bilingual dictionaries) and it is based on Latent Semantic Analysis (LSA) (Deerwester et al., 1990). MDMs are then exploited to define a Multilingual Domain Kernel, a generalized similarity function among documents in different languages that exploits a MDM (see Section 5). The Multilingual Domain Kernel is used inside a Support Vector Machines (SVM) classification framework for TC (Joachims, 2002). In Section 6 we will evaluate our technique in a Cross Language categorization task. The results show that our approach is a feasible and cheap solution, largely outperforming a baseline. Conclusions and future works are finally reported in Section 7.

2 Comparable Corpora

Comparable corpora are collections of texts in different languages regarding similar topics (e.g. a collection of news published by agencies in the same period). More restrictive requirements are expected for parallel corpora (i.e. corpora composed by texts which are mutual translations), while the class of the multilingual corpora (i.e. collection of texts expressed in different languages without any additional requirement) is the more general. Obviously parallel corpora are also comparable, while comparable corpora are also multilingual.

In a more precise way, let $L = \{L^1, L^2, \ldots, L^l\}$ be a set of languages, let $T^i = \{t_1^i, t_2^i, \ldots, t_n^i\}$ be a collection of texts expressed in the language $L^i \in L$, and let $\psi(t_h^j, t_z^i)$ be a function that returns 1 if t_z^i is the translation of t_h^j and 0 otherwise. A *multilingual corpus* is the collection of texts defined by $T^* = \bigcup_i T^i$. If the function ψ exists for every text $t_z^i \in T^*$ and for every language L^j, and is known, then the corpus is *parallel* and *aligned* at document level.

For the purpose of this paper it is enough to assume that two corpora are comparable, i.e. they are composed by documents about the same topics and produced in the same period (e.g. possibly from different news agencies), and it is not known if a function ψ exists, even if in principle it could exist and return 1 for a strict subset of document pairs.

There exist many interesting works about using parallel corpora for multilingual applications (Melamed, 2001), such as Machine Translation, Cross language Information Retrieval (Littman et al., 1998), lexical acquisition, and so on.

However it is not always easy to find or build parallel corpora. This is the main reason because the *weaker* notion of comparable corpora is a matter recent interest in the field of Computational Linguistics (Gaussier et al., 2004).

The texts inside comparable corpora, being about the same topics (i.e. about the same semantic domains), should refer to the same concepts by using various expressions in different languages. On the other hand, most of the proper nouns, relevant entities and words that are not yet lexicalized in the language, are expressed by using their original terms. As a consequence the *same entities* will be denoted with the *same words* in different languages, allowing to automatically detect couples of translation pairs just by looking at the word shape (Koehn and Knight, 2002). Our hypothesis is that comparable corpora contain a large amount of such words, just because texts, referring to the same topics in different languages, will often adopt the same terms to denote the same entities[1].

However, the simple presence of these shared words is not enough to get significant results in TC tasks. As we will see, we need to exploit these common words to induce a second-order similarity for the other words in the lexicons.

3 The Multilingual Vector Space Model

Let $T = \{t_1, t_2, \ldots, t_n\}$ be a corpus, and $V = \{w_1, w_2, \ldots, w_k\}$ be its vocabulary. In the monolingual settings, the Vector Space Model (VSM) is a k-dimensional space $\mathbf{R}^k$, in which the text $t_j \in T$ is represented by means of the vector $\vec{t_j}$ such that the z^{th} component of $\vec{t_j}$ is the frequency of w_z in t_j. The similarity among two texts in the VSM is then estimated by computing the cosine of their vectors in the VSM.

[1] According to our assumption, a possible additional criterion to decide whether two corpora are comparable is to estimate the percentage of terms in the intersection of their vocabularies.

Unfortunately, such a model cannot be adopted in the multilingual settings, because the VSMs of different languages are mainly disjoint, and the similarity between two texts in different languages would always turn out zero. This situation is represented in Figure 1, in which both the left-bottom and the rigth-upper regions of the matrix are totally filled by zeros.

A first attempt to solve this problem is to exploit the information provided by external knowledge sources, such as bilingual dictionaries, to collapse all the rows representing translation pairs. In this setting, the similarity among texts in different languages could be estimated by exploiting the classical VSM just described. However, the main disadvantage of this approach to estimate inter-lingual text similarity is that it strongly relies on the availability of a multilingual lexical resource containing a list of translation pairs. For languages with scarce resources a bilingual dictionary could be not easily available. Secondly, an important requirement of such a resource is its coverage (i.e. the amount of possible translation pairs that are actually contained in it). Finally, another problem is that ambiguos terms could be translated in different ways, leading to collapse together rows describing terms with very different meanings.

On the other hand, the assumption of corpora comparability seen in Section 2, implies the presence of a number of common words, represented by the central rows of the matrix in Figure 1.

As we will show in Section 6, this model is rather poor because of its sparseness. In the next section, we will show how to use such words as seeds to induce a Multilingual Domain VSM, in which second order relations among terms and documents in different languages are considered to improve the similarity estimation.

4 Multilingual Domain Models

A MDM is a multilingual extension of the concept of Domain Model. In the literature, Domain Models have been introduced to represent ambiguity and variability (Gliozzo et al., 2004) and successfully exploited in many NLP applications, such us Word Sense Disambiguation (Strapparava et al., 2004), Text Categorization and Term Categorization.

A Domain Model is composed by soft clusters of terms. Each cluster represents a semantic domain, i.e. a set of terms that often co-occur in texts having similar topics. Such clusters identifies groups of words belonging to the same semantic field, and thus highly paradigmatically related. MDMs are Domain Models containing terms in more than one language.

A MDM is represented by a matrix $\mathbf{D}$, containing the degree of association among terms in all the languages and domains, as illustrated in Table 1.

	MEDICINE	COMPUTER_SCIENCE
$HIV^{e/i}$	1	0
$AIDS^{e/i}$	1	0
$virus^{e/i}$	0.5	0.5
$hospital^{e}$	1	0
$laptop^{e}$	0	1
$Microsoft^{e/i}$	0	1
$clinica^{i}$	1	0

Table 1: Example of Domain Matrix. w^e denotes English terms, w^i Italian terms and $w^{e/i}$ the common terms to both languages.

MDMs can be used to describe lexical ambiguity, variability and inter-lingual domain relations. Lexical ambiguity is represented by associating one term to more than one domain, while variability is represented by associating different terms to the same domain. For example the term *virus* is associated to both the domain COMPUTER_SCIENCE and the domain MEDICINE while the domain MEDICINE is associated to both the terms *AIDS* and *HIV*. Inter-lingual domain relations are captured by placing different terms of different languages in the same semantic field (as for example $HIV^{e/i}$, $AIDS^{e/i}$, $hospital^{e}$, and $clinica^{i}$). Most of the named entities, such as *Microsoft* and *HIV* are expressed using the same string in both languages.

When similarity among texts in different languages has to be estimated, the information contained in the MDM is crucial. For example the two sentences *"I went to the hospital to make an HIV check"* and *"Ieri ho fatto il test dell'AIDS in clinica"* (lit. *yesterday I did the AIDS test in a clinic*) are very highly related, even if they share no tokens. Having an "a priori" knowledge about the inter-lingual domain similarity among *AIDS*, *HIV*, *hospital* and *clinica* is then a useful information to

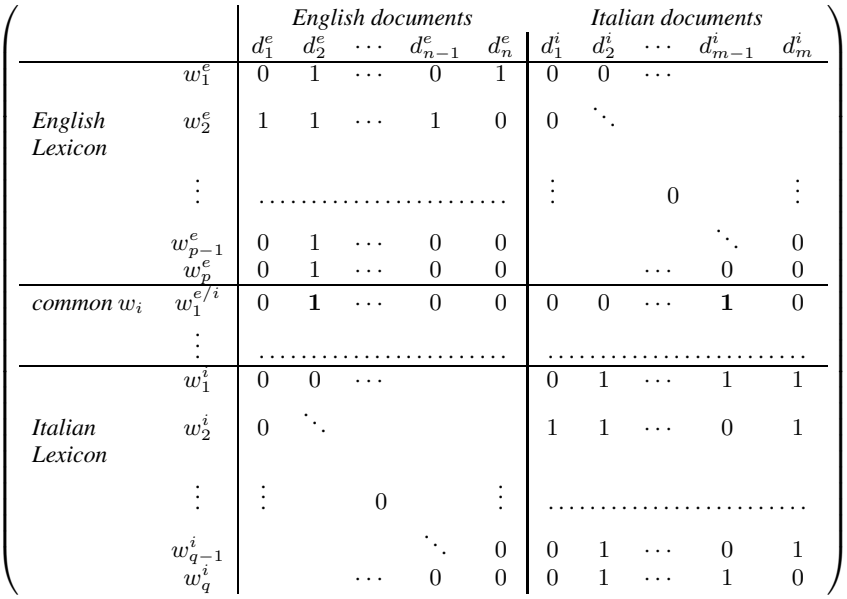

Figure 1: Multilingual term-by-document matrix

recognize inter-lingual topic similarity. Obviously this relation is less restrictive than a stronger association among translation pair. In this paper we will show that such a representation is sufficient for TC puposes, and easier to acquire.

In the rest of this section we will provide a formal definition of the concept of MDM, and we define some similarity metrics that exploit it.

Formally, let $V^i = \{w_1^i, w_2^i, \ldots, w_{k_i}^i\}$ be the vocabulary of the corpus T^i composed by document expressed in the language L^i, let $V^* = \bigcup_i V^i$ be the set of all the terms in all the languages, and let $k^* = |V^*|$ be the cardinality of this set. Let $\mathcal{D} = \{D_1, D_2, ..., D_d\}$ be a set of domains. A DM is fully defined by a $k^* \times d$ *domain matrix* $\mathbf{D}$ representing in each cell $\mathbf{d_{i,z}}$ the *domain relevance* of the i^{th} term of V^* with respect to the domain D_z. The domain matrix $\mathbf{D}$ is used to define a function $\mathcal{D} : \mathbf{R}^{k^*} \rightarrow \mathbf{R}^d$, that maps the document vectors $\vec{t_j}$ expressed into the multilingual classical VSM, into the vectors $\vec{t_j'}$ in the *multilingual domain VSM*. The function $\mathcal{D}$ is defined by[2]

$$\mathcal{D}(\vec{t_j}) = \vec{t_j}(\mathbf{I^{IDF}D}) = \vec{t_j'} \quad (1)$$

where $\mathbf{I^{IDF}}$ is a diagonal matrix such that $i_{i,i}^{IDF} = IDF(w_i^l)$, $\vec{t_j}$ is represented as a row vector, and $IDF(w_i^l)$ is the *Inverse Document Frequency* of w_i^l evaluated in the corpus T^l.

The matrix $\mathbf{D}$ can be determined for example using hand-made lexical resources, such as WORDNET DOMAINS (Magnini and Cavaglià, 2000). In the present work we followed the way to acquire $\mathbf{D}$ automatically from corpora, exploiting the technique described below.

4.1 Automatic Acquisition of Multilingual Domain Models

In this work we propose the use of Latent Semantic Analysis (LSA) (Deerwester et al., 1990) to induce a MDM from comparable corpora. LSA is an unsupervised technique for estimating the similarity among texts and terms in a large corpus. In the monolingual settings LSA is performed by means of a Singular Value Decomposition (SVD) of the term-by-document matrix $\mathbf{T}$ describing the corpus. SVD decomposes the term-by-document matrix $\mathbf{T}$ into three matrixes $\mathbf{T} \simeq \mathbf{V}\mathbf{\Sigma_{k'}}\mathbf{U}^T$ where $\mathbf{\Sigma_{k'}}$ is the diagonal $k \times k$ matrix containing the highest $k' \ll k$

[2]In (Wong et al., 1985) the formula 1 is used to define a Generalized Vector Space Model, of which the Domain VSM is a particular instance.

eigenvalues of $\mathbf{T}$, and all the remaining elements are set to 0. The parameter k' is the dimensionality of the Domain VSM and can be fixed in advance (i.e. $k' = d$).

In the literature (Littman et al., 1998) LSA has been used in multilingual settings to define a multilingual space in which texts in different languages can be represented and compared. In that work LSA strongly relied on the availability of aligned parallel corpora: documents in all the languages are represented in a term-by-document matrix (see Figure 1) and then the columns corresponding to sets of translated documents are collapsed (i.e. they are substituted by their sum) before starting the LSA process. The effect of this step is to merge the subspaces (i.e. the right and the left sectors of the matrix in Figure 1) in which the documents have been originally represented.

In this paper we propose a variation of this strategy, performing a multilingual LSA in the case in which an aligned parallel corpus is not available.

It exploits the presence of common words among different languages in the term-by-document matrix. The SVD process has the effect of creating a LSA space in which documents in both languages are represented. Of course, the higher the number of common words, the more information will be provided to the SVD algorithm to find common LSA dimension for the two languages. The resulting LSA dimensions can be perceived as multilingual clusters of terms and document. LSA can then be used to define a Multilingual Domain Matrix $\mathbf{D_{LSA}}$.

$$\mathbf{D_{LSA}} = \mathbf{I^N V}\sqrt{\mathbf{\Sigma_{k'}}} \quad (2)$$

where $\mathbf{I^N}$ is a diagonal matrix such that $\mathbf{i^N_{i,i}} = \frac{1}{\sqrt{\langle \vec{w'_i}, \vec{w'_i} \rangle}}$, $\vec{w'_i}$ is the i^{th} row of the matrix $\mathbf{V}\sqrt{\mathbf{\Sigma_{k'}}}$.

Thus $\mathbf{D_{LSA}}$[3] can be exploited to estimate similarity among texts expressed in different languages (see Section 5).

[3]When $\mathbf{D_{LSA}}$ is substituted in Equation 1 the Domain VSM is equivalent to a Latent Semantic Space (Deerwester et al., 1990). The only difference in our formulation is that the vectors representing the terms in the Domain VSM are normalized by the matrix $\mathbf{I^N}$, and then rescaled, according to their IDF value, by matrix $\mathbf{I^{IDF}}$. Note the analogy with the *tf idf* term weighting schema, widely adopted in Information Retrieval.

4.2 Similarity in the multilingual domain space

As an example of the second-order similarity provided by this approach, we can see in Table 2 the five most similar terms to the lemma *bank*. The similarity among terms is calculated by cosine among the rows in the matrix $\mathbf{D_{LSA}}$, acquired from the data set used in our experiments (see Section 6.2). It is worth noting that the Italian lemma *banca* (i.e. bank in English) has a high similarity score to the English lemma *bank*. While this is not enough to have a precise term translation, it is sufficient to capture relevant aspects of topic similarity in a cross-language text categorization task.

Lemma#Pos	*Similarity Score*	*Language*
banking#n	0.96	Eng
credit#n	0.90	Eng
amro#n	0.89	Eng
unicredito#n	0.85	Ita
banca#n	0.83	Ita

Table 2: Terms with high similarity to the English lemma *bank#n*, in the Multilingual Domain Model

5 The Multilingual Domain Kernel

Kernel Methods are the state-of-the-art supervised framework for learning, and they have been successfully adopted to approach the TC task (Joachims, 2002).

The basic idea behind kernel methods is to embed the data into a suitable feature space $\mathcal{F}$ via a mapping function $\phi : \mathcal{X} \rightarrow \mathcal{F}$, and then to use a linear algorithm for discovering nonlinear patterns. Kernel methods allow us to build a modular system, as the kernel function acts as an interface between the data and the learning algorithm. Thus the kernel function becomes the only domain specific module of the system, while the learning algorithm is a general purpose component. Potentially any kernel function can work with any kernel-based algorithm, as for example Support Vector Machines (SVMs).

During the learning phase SVMs assign a weight $\lambda_i \geq 0$ to any example $x_i \in X$. All the labeled instances x_i such that $\lambda_i > 0$ are called Support Vectors. Support Vectors lie close to the best separating hyper-plane between positive and negative examples. New examples are then assigned to the class

of the closest support vectors, according to equation 3.

$$f(x) = \sum_{i=1}^{n} \lambda_i K(x_i, x) + \lambda_0 \quad (3)$$

The kernel function $K(x_i, x)$ returns the similarity between two instances in the input space X, and can be designed just by taking care that some formal requirements are satisfied, as described in (Schölkopf and Smola, 2001).

In this section we define the Multilingual Domain Kernel, and we apply it to a cross language TC task. This kernel can be exploited to estimate the topic similarity among two texts expressed in different languages by taking into account the external knowledge provided by a MDM. It defines an explicit mapping $\mathcal{D} : \mathbf{R}^k \rightarrow \mathbf{R}^{k'}$ from the Multilingual VSM into the Multilingual Domain VSM. The Multilingual Domain Kernel is specified by

$$K_D(t_i, t_j) = \frac{\langle \mathcal{D}(t_i), \mathcal{D}(t_j) \rangle}{\sqrt{\langle \mathcal{D}(t_j), \mathcal{D}(t_j) \rangle \langle \mathcal{D}(t_i), \mathcal{D}(t_i) \rangle}} \quad (4)$$

where $\mathcal{D}$ is the Domain Mapping defined in equation 1. Thus the Multilingual Domain Kernel requires Multilingual Domain Matrix $\mathbf{D}$, in particular $\mathbf{D_{LSA}}$ that can be acquired from comparable corpora, as explained in Section 4.1.

To evaluate the Multilingual Domain Kernel we compared it to a baseline kernel function, namely the *bag_of_words* kernel, that simply estimates the topic similarity in the Multilingual VSM, as described in Section 3. The BoW kernel is a particular case of the Domain Kernel, in which $\mathbf{D} = \mathbf{I}$, and $\mathbf{I}$ is the identity matrix.

6 Evaluation

In this section we present the data set (two comparable English and Italian corpora) used in the evaluation, and we show the results of the Cross Language TC tasks. In particular we tried both to train the system on the English data set and classify Italian documents and to train using Italian and classify the English test set. We compare the learning curves of the Multilingual Domain Kernel with the standard BoW kernel, which is considered as a baseline for this task.

6.1 Implementation details

As a supervised learning device, we used the SVM implementation described in (Joachims, 1999). The Multilingual Domain Kernel is implemented by defining an explicit feature mapping as explained above, and by normalizing each vector. All the experiments have been performed with the standard SVM parameter settings.

We acquired a Multilingual Domain Model by performing the Singular Value Decomposition process on the term-by-document matrices representing the merged training partitions (i.e. English and Italian), and we considered only the first 400 dimensions[4].

6.2 Data set description

We used a news corpus kindly put at our disposal by ADNKRONOS, an important Italian news provider. The corpus consists of 32,354 Italian and 27,821 English news partitioned by ADNKRONOS in a number of four fixed categories: `Quality_of_Life`, `Made_in_Italy`, `Tourism`, `Culture_and_School`. The corpus is comparable, in the sense stated in Section 2, i.e. they covered the same topics and the same period of time. Some news are translated in the other language (but *no* alignment indication is given), some others are present only in the English set, and some others only in the Italian. The average length of the news is about 300 words. We randomly split both the English and Italian part into 75% training and 25% test (see Table 3). In both the data sets we postagged the texts and we considered only the noun, verb, adjective, and adverb parts of speech, representing them by vectors containing the frequencies of each lemma with its part of speech.

6.3 Monolingual Results

Before going to a cross-language TC task, we conducted two tests of classical monolingual TC by training and testing the system on Italian and English documents separately. For these tests we used the SVM with the BoW kernel. Figures 2 and 3 report the results.

[4]To perform the SVD operation we used LIBSVDC `http://tedlab.mit.edu/~dr/SVDLIBC/`.

Categories	*English* Training	Test	Total	*Italian* Training	Test	Total
Quality_of_Life	5759	1989	7748	5781	1901	7682
Made_in_Italy	5711	1864	7575	6111	2068	8179
Tourism	5731	1857	7588	6090	2015	8105
Culture_and_School	3665	1245	4910	6284	2104	8388
Total	20866	6955	27821	24266	8088	32354

Table 3: Number of documents in the data set partitions

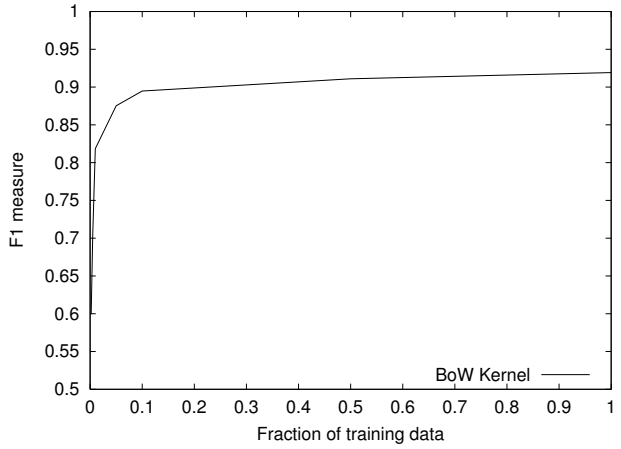

Figure 2: Learning curves for the English part of the corpus

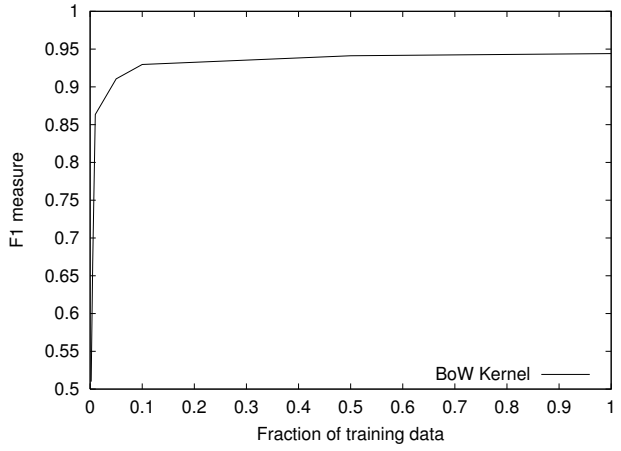

Figure 3: Learning curves for the Italian part of the corpus

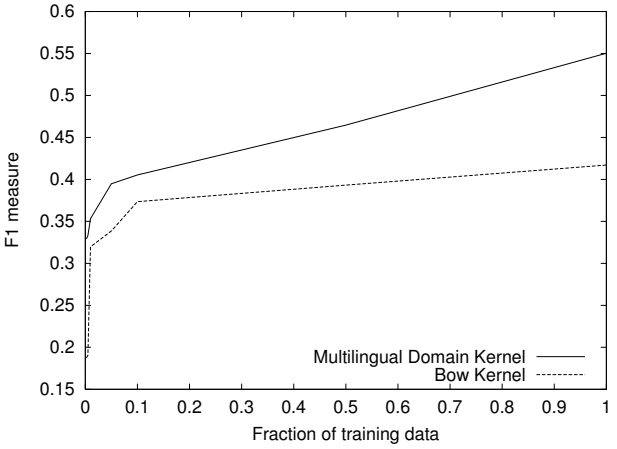

Figure 4: Cross-language (training on Italian, test on English) learning curves

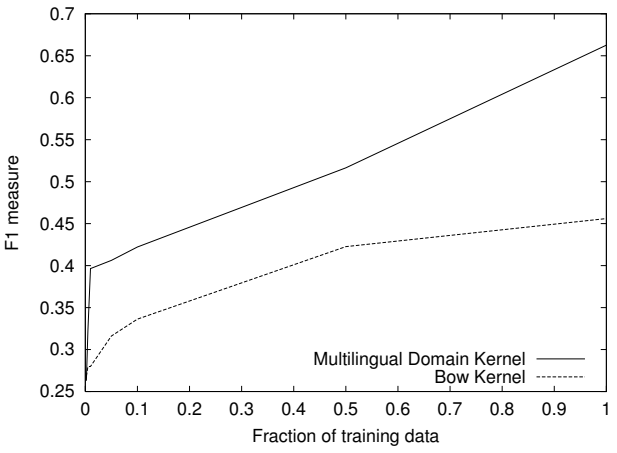

Figure 5: Cross-language (training on English, test on Italian) learning curves

6.4 A Cross Language Text Categorization task

As far as the cross language TC task is concerned, we tried the two possible options: we trained on the English part and we classified the Italian part, and we trained on the Italian and classified on the English part. The Multilingual Domain Model was acquired running the SVD only on the joint (English and Italian) training parts.

Table 4 reports the vocabulary dimensions of the English and Italian training partitions, the vocabu-

	# lemmata
English training	22,704
Italian training	26,404
English + Italian	43,384
common lemmata	5,724

Table 4: Number of lemmata in the training parts of the corpus

lary of the merged training, and how many common lemmata are present (about 14% of the total). Among the common lemmata, 97% are nouns and most of them are proper nouns. Thus the initial term-by-document matrix is a 43,384 $\times$ 45,132 matrix, while the $\mathbf{D_{LSA}}$ matrix is 43,384 $\times$ 400. For this task we consider as a baseline the BoW kernel.

The results are reported in Figures 4 and 5. Analyzing the learning curves, it is worth noting that when the quantity of training increases, the performance becomes better and better for the Multilingual Domain Kernel, suggesting that with more available training it could be possible to go closer to typical monolingual TC results.

7 Conclusion

In this paper we proposed a solution to cross language Text Categorization based on acquiring Multilingual Domain Models from comparable corpora in a totally unsupervised way and without using any external knowledge source (e.g. bilingual dictionaries). These Multilingual Domain Models are exploited to define a generalized similarity function (i.e. a kernel function) among documents in different languages, which is used inside a Support Vector Machines classification framework. The basis of the similarity function exploits the presence of common words to induce a second-order similarity for the other words in the lexicons. The results have shown that this technique is sufficient to capture relevant aspects of topic similarity in cross-language TC tasks, obtaining substantial improvements over a simple baseline. As future work we will investigate the performance of this approach to more than two languages TC task, and a possible generalization of the assumption about equality of the common words.

Acknowledgments

This work has been partially supported by the ONTOTEXT project, funded by the Autonomous Province of Trento under the FUP-2004 program.

References

S. Deerwester, S. T. Dumais, G. W. Furnas, T.K. Landauer, and R. Harshman. 1990. Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41(6):391–407.

E. Gaussier, J. M. Renders, I. Matveeva, C. Goutte, and H. Dejean. 2004. A geometric view on bilingual lexicon extraction from comparable corpora. In *Proceedings of ACL-04*, Barcelona, Spain, July.

A. Gliozzo, C. Strapparava, and I. Dagan. 2004. Unsupervised and supervised exploitation of semantic domains in lexical disambiguation. *Computer Speech and Language*, 18:275–299.

T. Joachims. 1999. Making large-scale SVM learning practical. In B. Schölkopf, C. Burges, and A. Smola, editors, *Advances in kernel methods: support vector learning*, chapter 11, pages 169 – 184. The MIT Press.

T. Joachims. 2002. *Learning to Classify Text using Support Vector Machines*. Kluwer Academic Publishers.

P. Koehn and K. Knight. 2002. Learning a translation lexicon from monolingual corpora. In *Proceedings of ACL Workshop on Unsupervised Lexical Acquisition*, Philadelphia, July.

M. Littman, S. Dumais, and T. Landauer. 1998. Automatic cross-language information retrieval using latent semantic indexing. In G. Grefenstette, editor, *Cross Language Information Retrieval*, pages 51–62. Kluwer Academic Publishers.

B. Magnini and G. Cavaglià. 2000. Integrating subject field codes into WordNet. In *Proceedings of LREC-2000*, Athens, Greece, June.

D. Melamed. 2001. *Empirical Methods for Exploiting Parallel Texts*. The MIT Press.

B. Schölkopf and A. J. Smola. 2001. *Learning with Kernels. Support Vector Machines, Regularization, Optimization, and Beyond*. The MIT Press.

C. Strapparava, A. Gliozzo, and C. Giuliano. 2004. Pattern abstraction and term similarity for word sense disambiguation. In *Proceedings of SENSEVAL-3*, Barcelona, Spain, July.

S.K.M. Wong, W. Ziarko, and P.C.N. Wong. 1985. Generalized vector space model in information retrieval. In *Proceedings of the 8^{th} ACM SIGIR Conference*.