Exploiting Comparable Corpora and Bilingual Dictionaries for Cross-Language Text Categorization

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

Abstract

Cross-language Text Categorization is the task of assigning semantic classes to documents written in a target language (e.g. English) while the system is trained using labeled documents in a source language (e.g. Italian).

In this work we present many solutions according to the availability of bilingual resources, and we show that it is possible to deal with the problem even when no such resources are accessible. The core technique relies on the automatic acquisition of Multilingual Domain Models from comparable corpora.

Experiments show the effectiveness of our approach, providing a low cost solution for the Cross Language Text Categorization task. In particular, when bilingual dictionaries are available the performance of the categorization gets close to that of monolingual text categorization.

1 Introduction

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 Natural Language Processing (NLP) problems into a multilingual setting. For instance the classical monolingual Text Categorization (TC) problem can be reformulated as a *Cross Language Text Categorization* (CLTC) task, in which 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*).

The applicative interest for the CLTC is immediately clear in the globalized Web scenario. For example, in the community based trade (e.g. eBay) it is often necessary to archive texts in different languages by adopting common merceological categories, very often defined by collections of documents in a source language (e.g. English). Another application along this direction is Cross Lingual Question Answering, in which it would be very useful to filter out the candidate answers according to their topics.

In the literature, this task has been proposed quite recently (Bel et al., 2003; Gliozzo and Strapparava, 2005). In those works, authors exploited comparable corpora showing promising results. A more recent work (Rigutini et al., 2005) proposed the use of Machine Translation techniques to approach the same task.

Classical approaches for multilingual problems have been conceived by following two main directions: (i) knowledge based approaches, mostly implemented by rule based systems and (ii) empirical approaches, in general relying on statistical learning from parallel corpora. Knowledge based approaches are often affected by low accuracy. Such limitation is mainly due to the problem of tuning large scale multilingual lexical resources (e.g. MultiWordNet, EuroWordNet) for the specific application task (e.g. discarding irrelevant senses, extending the lexicon with domain specific terms and their translations). On the other hand, empirical approaches are in general more accurate, because they can be trained from domain specific collections of parallel text to represent the application needs. There exist many interesting works about using parallel corpora for multilingual applications (Melamed, 2001), such as Machine Translation (Callison-Burch et al., 2004), Cross Lingual

Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the ACL, pages 553–560,
Sydney, July 2006. ©2006 Association for Computational Linguistics

Information Retrieval (Littman et al., 1998), and so on.

However it is not always easy to find or build parallel corpora. This is the main reason why the "weaker" notion of comparable corpora is a matter of recent interest in the field of Computational Linguistics (Gaussier et al., 2004). In fact, comparable corpora are easier to collect for most languages (e.g. collections of international news agencies), providing a low cost knowledge source for multilingual applications.

The main problem of adopting comparable corpora for multilingual knowledge acquisition is that only weaker statistical evidence can be captured. In fact, while parallel corpora provide stronger (text-based) statistical evidence to detect translation pairs by analyzing term co-occurrences in translated documents, comparable corpora provides weaker (term-based) evidence, because text alignments are not available.

In this paper we present some solutions to deal with CLTC according to the availability of bilingual resources, and we show that it is possible to deal with the problem even when no such resources are accessible. The core technique relies on the automatic acquisition of Multilingual Domain Models (MDMs) from comparable corpora. This allows us to define a kernel function (i.e. a similarity function among documents in different languages) that is then exploited inside a Support Vector Machines classification framework. We also investigate this problem exploiting synset-aligned multilingual WordNets and standard bilingual dictionaries (e.g. Collins).

Experiments show the effectiveness of our approach, providing a simple and low cost solution for the Cross-Language Text Categorization task. In particular, when bilingual dictionaries/repositories are available, the performance of the categorization gets close to that of monolingual TC.

The paper is structured as follows. Section 2 briefly discusses the notion of comparable corpora. Section 3 shows how to perform cross-lingual TC when no bilingual dictionaries are available and it is possible to rely on a comparability assumption. Section 4 present a more elaborated technique to acquire MDMs exploiting bilingual resources, such as MultiWordNet (i.e. a synset-aligned WordNet) and Collins bilingual dictionary. Section 5 evaluates our methodologies and Section 6 concludes the paper suggesting some future developments.

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 of 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 $T^i = \{t^i_1, t^i_2, \ldots, t^i_n\}$ be a collection of texts expressed in the language $L^i \in L$, and let $\psi(t^j_h, t^i_z)$ be a function that returns 1 if t^i_z is the translation of t^j_h 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 $t^i_z \in T^*$ 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 of 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.

The texts inside comparable corpora, being about the same topics, 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 us 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].

[1] According to our assumption, a possible additional cri-

However, the simple presence of these shared words is not enough to get significant results in CLTC 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.

2.1 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.

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 to be 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.

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 5, 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.

3 Exploiting Comparable Corpora

Looking at the multilingual term-by-document matrix in Figure 1, a first attempt to merge the subspaces associated to each language is to exploit the information provided by external knowledge sources, such as bilingual dictionaries, e.g. collapsing 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. 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 ambiguous terms could be translated in different ways, leading us to collapse together rows describing terms with very different meanings. In Section 4 we will see how the availability of bilingual dictionaries influences the techniques and the performance. In the present Section we want to explore the case in which such resources are supposed not available.

3.1 Multilingual Domain Model

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 as Word Sense Disambiguation (Strapparava et al., 2004), Text Categorization and Term Categorization.

A Domain Model is composed of 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 identify 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. 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

	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.

terion to decide whether two corpora are comparable is to estimate the percentage of terms in the intersection of their vocabularies.

$$\left(\begin{array}{ll|ccccc|ccccc}
 & & \multicolumn{5}{c|}{\textit{English texts}} & \multicolumn{5}{c}{\textit{Italian texts}} \\
 & & t_1^e & t_2^e & \cdots & t_{n-1}^e & t_n^e & t_1^i & t_2^i & \cdots & t_{m-1}^i & t_m^i \\
\hline
 & w_1^e & 0 & 1 & \cdots & 0 & 1 & 0 & 0 & \cdots & & \\
\textit{English} & w_2^e & 1 & 1 & \cdots & 1 & 0 & 0 & \ddots & & & \\
\textit{Lexicon} & & & & & & & & & & & \\
 & \vdots & \multicolumn{5}{c|}{\dots\dots\dots\dots\dots\dots\dots} & \vdots & & 0 & & \vdots \\
 & w_{p-1}^e & 0 & 1 & \cdots & 0 & 0 & & & & \ddots & 0 \\
 & w_p^e & 0 & 1 & \cdots & 0 & 0 & & & \cdots & 0 & 0 \\
\hline
\textit{common}\ w_i & w_1^{e/i} & 0 & \mathbf{1} & \cdots & 0 & 0 & 0 & 0 & \cdots & \mathbf{1} & 0 \\
 & \vdots & \multicolumn{5}{c|}{\dots\dots\dots\dots\dots\dots\dots} & \multicolumn{5}{c}{\dots\dots\dots\dots\dots\dots\dots} \\
\hline
 & w_1^i & 0 & 0 & \cdots & & & 0 & 1 & \cdots & 1 & 1 \\
\textit{Italian} & w_2^i & 0 & \ddots & & & & 1 & 1 & \cdots & 0 & 1 \\
\textit{Lexicon} & & & & & & & & & & & \\
 & \vdots & \vdots & & 0 & & \vdots & \multicolumn{5}{c}{\dots\dots\dots\dots\dots\dots\dots} \\
 & w_{q-1}^i & & & & \ddots & 0 & 0 & 1 & \cdots & 0 & 1 \\
 & w_q^i & & & \cdots & 0 & 0 & 0 & 1 & \cdots & 1 & 0
\end{array}\right)$$

Figure 1: Multilingual term-by-document matrix

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.

Formally, let $V^i = \{w_1^i, w_2^i, \dots, w_{k_i}^i\}$ be the vocabulary of the corpus T^i composed of 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 (see Section 2.1), into the vectors $\vec{t'_j}$ in the *multilingual domain VSM*. The function $\mathcal{D}$ is defined by[2]

$$\mathcal{D}(\vec{t_j}) = \vec{t_j}(\mathbf{I^{IDF}D}) = \vec{t'_j} \qquad (1)$$

where $\mathbf{I^{IDF}}$ is a diagonal matrix such that $i_{i,l}^{IDF} = IDF(w_i^l)$, $\vec{t_j}$ 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.

In this work we exploit Latent Semantic Analysis (LSA) (Deerwester et al., 1990) to automatically acquire 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 $\mathbf{T} \simeq \mathbf{V\Sigma_{k'}U}^T$ where $\mathbf{\Sigma_{k'}}$ is the diagonal $k \times k$ matrix containing the highest $k' \ll k$ 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

[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.

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 $\mathbf{D_{LSA}}$. For further details see (Gliozzo and Strapparava, 2005).

As 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), we chose this framework to perform all our experiments, in particular Support Vector Machines[3]. Taking into account the external knowledge provided by a MDM it is possible estimate the topic similarity among two texts expressed in different languages, with the following kernel:

$$K_D(t_i, t_j) = \frac{\langle \mathcal{D}(t_i), \mathcal{D}(t_j) \rangle}{\sqrt{\langle \mathcal{D}(t_j), \mathcal{D}(t_j) \rangle \langle \mathcal{D}(t_i), \mathcal{D}(t_i) \rangle}} \quad (2)$$

where $\mathcal{D}$ is defined as in equation 1.

Note that when we want to estimate the similarity in the standard Multilingual VSM, as described in Section 2.1, we can use a simple *bag_of_words* kernel. The BoW kernel is a particular case of the Domain Kernel, in which $\mathbf{D} = \mathbf{I}$, and $\mathbf{I}$ is the identity matrix. In the evaluation typically we consider the BoW Kernel as a baseline.

4 Exploiting Bilingual Dictionaries

When bilingual resources are available it is possible to augment the the "common" portion of the matrix in Figure 1. In our experiments we exploit two alternative multilingual resources: MultiWordNet and the Collins English-Italian bilingual dictionary.

MultiWordNet[4]. It is a multilingual computational lexicon, conceived to be strictly aligned with the Princeton WordNet. The available languages are Italian, Spanish, Hebrew and Romanian. In our experiment we used the English and the Italian components. The last version of the Italian WordNet contains around 58,000 Italian word senses and 41,500 lemmas organized into 32,700 synsets aligned whenever possible with WordNet English synsets. The Italian synsets are created in correspondence with the Princeton WordNet synsets, whenever possible, and semantic relations are imported from the corresponding English synsets. This implies that the synset index structure is the same for the two languages.

Thus for the all the monosemic words, we augment each text in the dataset with the corresponding *synset-id*, which act as an expansion of the "common" terms of the matrix in Figure 1. Adopting the methodology described in Section 3.1, we exploit these common sense-indexing to induce a second-order similarity for the other terms in the lexicons. We evaluate the performance of the cross-lingual text categorization, using both the BoW Kernel and the Multilingual Domain Kernel, observing that also in this case the leverage of the external knowledge brought by the MDM is effective.

It is also possible to augment each text with all the synset-ids of all the words (i.e. monosemic and polysemic) present in the dataset, hoping that the SVM machine learning device cut off the noise due to the inevitable *spurious* senses introduced in the training examples. Obviously in this case, differently from the "monosemic" enrichment seen above, it does not make sense to apply any dimensionality reduction supplied by the Multilingual Domain Model (i.e. the resulting second-order relations among terms and documents produced on a such "extended" corpus should not be meaningful)[5].

Collins. The Collins machine-readable bilingual dictionary is a medium size dictionary including 37,727 headwords in the English Section and 32,602 headwords in the Italian Section.

This is a traditional dictionary, without sense indexing like the WordNet repository. In this case

[3] We adopted the efficient implementation freely available at *http://svmlight.joachims.org/*.

[4] Available at *http://multiwordnet.itc.it*.

[5] The use of a WSD system would help in this issue. However the rationale of this paper is to see how far it is possible to go with very few resources. And we suppose that a multilingual all-words WSD system is not easily available.

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 2: Number of documents in the data set partitions

we follow the way, for each text of one language, to augment all the present words with the translation words found in the dictionary. For the same reason, we chose not to exploit the MDM, while experimenting along this way.

5 Evaluation

The CLTC task has been rarely attempted in the literature, and standard evaluation benchmark are not available. For this reason, we developed an evaluation task by adopting 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 into four fixed categories: QUALITY_OF_LIFE, MADE_IN_ITALY, TOURISM, CULTURE_AND_SCHOOL. The English and the Italian corpora are comparable, in the sense stated in Section 2, i.e. they cover the same topics and the same period of time. Some news stories 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 stories is about 300 words. We randomly split both the English and Italian part into 75% training and 25% test (see Table 2). We processed the corpus with PoS taggers, keeping only nouns, verbs, adjectives and adverbs.

Table 3 reports the vocabulary dimensions of the English and Italian training partitions, the vocabulary 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 × 45,132 matrix, while the $\mathbf{D_{LSA}}$ was acquired using 400 dimensions.

As far as the CLTC task is concerned, we tried the many possible options. In all the cases we trained on the English part and we classified the Italian part, and we trained on the Italian and classified on the English part. When used, the MDM was acquired running the SVD only on the joint (English and Italian) training parts.

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

Table 3: Number of lemmata in the training parts of the corpus

Using only comparable corpora. Figure 2 reports the performance without any use of bilingual dictionaries. Each graph show the learning curves respectively using a BoW kernel (that is considered here as a baseline) and the multilingual domain kernel. We can observe that the latter largely outperform a standard BoW approach. 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 improve the results.

Using bilingual dictionaries. Figure 3 reports the learning curves exploiting the addition of the synset-ids of the monosemic words in the corpus. As expected the use of a multilingual repository improves the classification results. Note that the MDM outperforms the BoW kernel.

Figure 4 shows the results adding in the English and Italian parts of the corpus all the synset-ids (i.e. monosemic and polisemic) and all the translations found in the Collins dictionary respectively. These are the best results we get in our experiments. In these figures we report also the performance of the corresponding monolingual TC (we used the SVM with the BoW kernel), which can be considered as an upper bound. We can observe that the CLTC results are quite close to the performance obtained in the monolingual classification tasks.

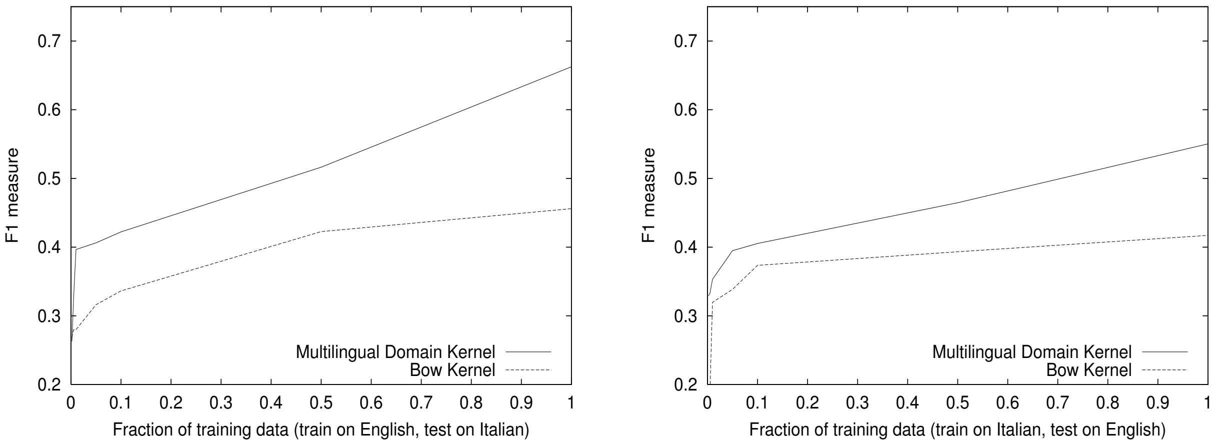

Figure 2: Cross-language learning curves: no use of bilingual dictionaries

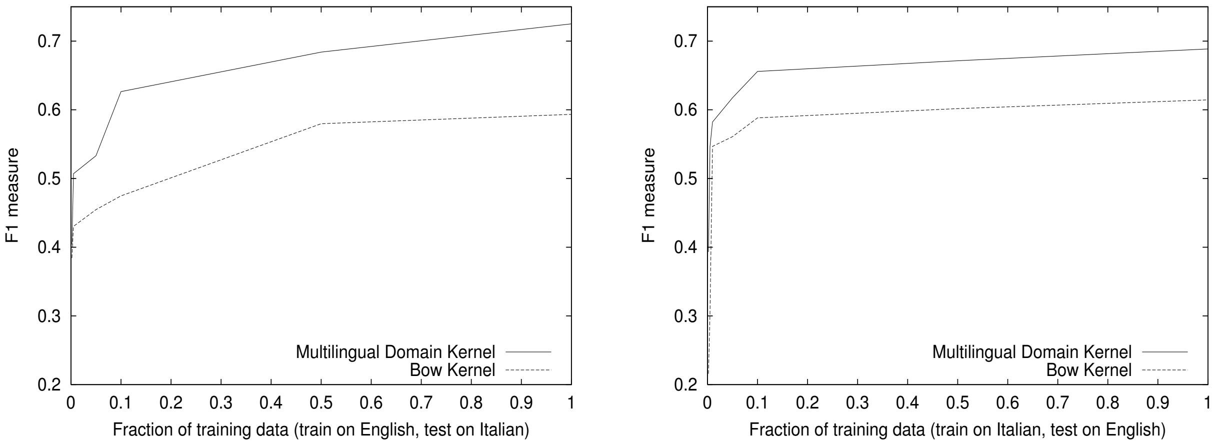

Figure 3: Cross-language learning curves: monosemic synsets from MultiWordNet

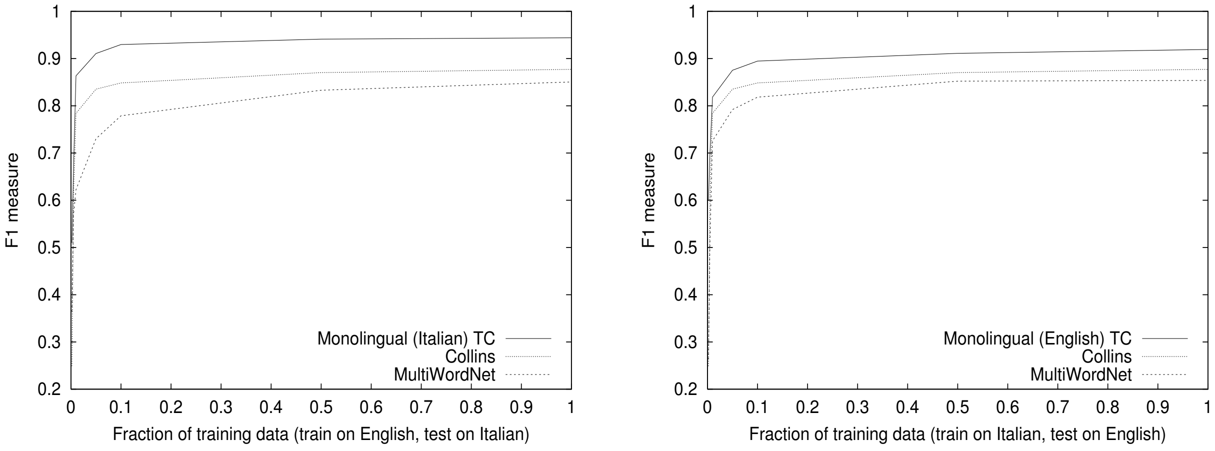

Figure 4: Cross-language learning curves: all synsets from MultiWordNet // All translations from Collins

6 Conclusion and Future Work

In this paper we have shown that the problem of cross-language text categorization on comparable corpora is a feasible task. In particular, it is possible to deal with it even when no bilingual resources are available. On the other hand when it is possible to exploit bilingual repositories, such as a synset-aligned WordNet or a bilingual dictionary, the obtained performance is close to that achieved for the monolingual task. In any case we think that our methodology is low-cost and simple, and it can represent a technologically viable solution for multilingual problems. For the future we try to explore also the use of a word sense disambiguation all-words system. We are confident that even with the actual state-of-the-art WSD performance, we can improve the actual results.

Acknowledgments

This work has been partially supported by the ONTOTEXT (From Text to Knowledge for the Semantic Web) project, funded by the Autonomous Province of Trento under the FUP-2004 program.

References

N. Bel, C. Koster, and M. Villegas. 2003. Cross-lingual text categorization. In *Proceedings of European Conference on Digital Libraries (ECDL)*, Trondheim, August.

C. Callison-Burch, D. Talbot, and M. Osborne. 2004. Statistical machine translation with word-and sentence-aligned parallel corpora. In *Proceedings of ACL-04*, Barcelona, Spain, July.

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 and C. Strapparava. 2005. Cross language text categorization by acquiring multilingual domain models from comparable corpora. In *Proc. of the ACL Workshop on Building and Using Parallel Texts (in conjunction of ACL-05)*, University of Michigan, Ann Arbor, June.

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. 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.

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

L. Rigutini, M. Maggini, and B. Liu. 2005. An EM based training algorithm for cross-language text categorizaton. In *Proceedings of Web Intelligence Conference (WI-2005)*, Compiègne, France, September.

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*.