Using linguistic information to classify Portuguese text documents

Teresa Gonçalves
Departamento de Informática
Universidade de Évora
7000-671 Évora, Portugal
tcg@di.uevora.pt

Paulo Quaresma
Departamento de Informática
Universidade de Évora
7000-671 Évora, Portugal
pq@di.uevora.pt

Abstract

This paper examines the role of various linguistic structures on text classification applying the study to the Portuguese language. Besides using a bag-of-words representation where we evaluate different measures and use linguistic knowledge for term selection, we do several experiments using syntactic information representing documents as strings of words and strings of syntactic parse trees.

To build the classifier we use the Support Vector Machine (SVM) algorithm which is known to produce good results on text classification tasks and apply the study to a dataset of articles from the Público newspaper. The results show that sentences' syntactic structure is not useful for text classification (as initially expected), but part-of-speech information can be used as a term selection technique to construct the bag-of-words representation of documents.

1. Introduction

Current Information Technologies and Web-based services need to manage, select and filter increasing amounts of textual information. Text classification allows users, through navigation on class hierarchies, to browse more easily the texts of their interests. This paradigm is very effective both in filtering information as in the development of online end-user services.

As the number of documents involved in these applications is large, efficient and automatic approaches are necessary for classification. Standard Machine Learning approaches use the bag-of-words representation to deceive the classification target function, where the only features are document word statistics. Typical linguistic structures such as morphology and syntax are usually neglected in the learning process. Moreover, almost all studies have been conducted on texts written in the English language.

In order to evaluate common Information Retrieval representation techniques (that possibly use some morphological information) for the Portuguese language, a set of initial experiments was made. The most effective document representation served as a starting point for the other analyses.

Then, the use of morphosyntactic and syntactic information was appreciated. While using morphosyntactic information as a term selecting technique, a structured representation based on parse trees was used to represent the syntactic information. To build the learners we used the Support Vector Machine (SVM) algorithm since it can support structured representations and is known to produce good results on text classification tasks.

This paper is organized as follows: Sections 2 presents concepts and tools related to linguistic information and Machine Learning, Section 3 describes the used document representation and Section 4 presents and Section 5 evaluates the experiments. Conclusions and future work are pointed out in Section 6.

2. Linguistic Information and Machine Learning

This section introduces the used linguistic concepts and tools and the chosen Machine Learning algorithm and software.

2.1. Linguistic information

The Portuguese language is morphological rich: while nouns and adjectives have 4 forms (two *genders* – masculine and feminine and two *numbers* – singular and plural), a regular verb has 66 different forms (two *numbers*, three *persons* – 1st, 2nd and 3rd and five *modes* – indicative, conjunctive, conditional, imperative and infinitive, each with different number of *tenses* ranging from 1 to 5).

2.1.1 Representation

Morphological information includes word stem and its morphological features, like grammatical class and flexion.

While some natural language processing tasks use word stem, others use its lemma.

Most syntactic language representations are based on the context-free grammar (CFG) formalism introduced by [4] and, independently, by [1]: given a sentence, it generates the corresponding syntactic structure usually represented through a tree structure known as sentence *parse tree*. It contains its constituents structure (such as noun and verb phrases) and words' grammatical class.

2.1.2 Tools

We applied a Portuguese stop-list (set of non-relevant words such as articles, pronouns, adverbs and prepositions) and POLARIS, a lexical database [9], to generate the lemma for each Portuguese word.

PALAVRAS [2] parser was used to obtain words' POS tags and sentences' parse tree. It was developed in the context of the VISL project by the Institute of Language and Communication of the University of Southern Denmark. This parser is robust enough to always produce an output even for incomplete or incorrect sentences and has a comparatively low percentage of errors (less than 1% for word class and 3-4% for surface syntax) [3].

Its output is the syntactic analysis of each phrase including the POS tag associated with each word. Possible POS tags and are: adjective (`adj`), adverb (`adv`), article (`det`), conjunction (`conj`), interjection (`in`), noun (`n`), numeral (`num`), preposition (`prp`), pronoun (`pron`), proper noun (`prop`) and verb (`v`).

2.2. Learning Algorithm

To build the learners we used the Support Vector Machine (SVM) algorithm (with different kernel functions) since it can support structured representations and is known to produce good results on text classification tasks.

2.2.1 Support Vector Machine

Support Vector Machine (SVM) is a learning algorithm introduced by Vapnik and coworkers [16], which was motivated by the theoretical results from the statistical learning theory. It joins a kernel technique with the structural risk minimization framework. A *kernel technique* comprises two parts: a module that performs a mapping into a suitable feature space and a learning algorithm designed to discover linear patterns in that space.

The *kernel function* (or simply the kernel), that implicitly performs the mapping, depends on the specific type and domain knowledge of the data source. The *learning algorithm* is general purpose and robust; it's also efficient, since the amount of computational resources required is polynomial with the size and number of data items, even when the dimension of the embedding space grows exponentially [15]. Its key aspects can be highlighted as follows (see Figure 1):

- Data items are embedded into a vector space called the feature space.
- Linear relations are discovered among images of data items in feature space.
- Algorithm is implemented in a way that the coordinates of the embedded points are not needed; only their pairwise inner products.
- Pairwise inner products can be computed efficiently directly from the original data using the kernel function.

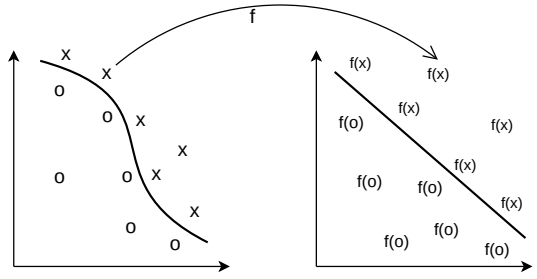

Figure 1. Kernel function: data nonlinear pattern transformed into linear feature space.

The *structural risk minimization* framework creates a model with a minimized VC (Vapnik-Chervonenkis) dimension. This developed theory [17] shows that when a model VC dimension is low, the expected probability of error is also low, which means good performance on unseen data.

2.2.2 Kernel functions

Most approaches to text classification use the basic vector space model (VSM) to represent documents. The simplest measure that takes into account word frequency in each document can be naturally reinterpreted as a kernel method. Normalization and term reduction approaches can also be interpreted as kernel functions (see [15]) and other standard kernels (like the polynomial one) apply non linear transformations to the usual VSM approach.

The *convolution kernel* [7] is the most well-known kernel for structured objects. A structured object is an object formed by the composition of simpler components; frequently, these components are, recursively, simpler objects of the same type. It's the case of strings, trees or graphs. The convolution kernel definition is based on kernels defined over structure components.

For tree structured objects, the feature space is indexed by subtrees and similarity is based on counting common subtrees. The subset tree kernel [5] is one of such kernels that uses ordered labelled trees counting subsets of common trees between two trees. They have produced good results on parse tree ranking [5] and predicate argument classification [11], [20].

2.2.3 Software

For the morphological and morphosyntactic levels we used WEKA SVM algorithm. WEKA [19] is a software package developed in New Zealand Waikato University implementing a large collection of ML algorithms.

For the syntactic level, the SVM algorithm was run using SVM^{light}-TK [10]. This software is an extension to SVM^{light} [8], that uses convolution kernels to represent tree structures. It implements two different kernels for tree structures: the subtree kernel [18] and the subset tree kernel [5]. Intuitively, the first counts all common n-descendants until the leaves (being n the root node) and the second adds to that counting all trees considering as leaves all internal tree nodes.

3. Document representation

Most text classification research is applied to English written documents and even if there are some that work on the sub-word or multi-word levels, the most used indexing term is, without doubt, the word. Next, we present the document representations used in this work.

3.1. Morphological information

As a baseline for this study, we first considered the traditional bag-of-words representation using the word and its lemma as indexing terms. We considered several filtering and weighting measures. We also used word grammatical class (morphosyntactic information) as a term selecting method. Figure 2 illustrates a two sentence document ("Mother observes her daughter Carlota. Carlota plays with the doll.") and the corresponding representation.

3.2. Syntactic information

Since a parse tree is an ordered tree, each document, that is a sequence of sentences, can be represented as an ordered tree of ordered trees. In this way, a document is a tree where each root child is the parse tree of a sentence and the leaves are its word lemma. This representation was named *syntactic-tree* representation.

However, we did not used the complete parse tree, but only the nodes of the following word classes: noun, proper noun, adjective, verb, pronoun, and adverb. Figure 3 illustrates a two sentence document ("Mother observes her daughter Carlota. Carlota plays with the doll.") and the corresponding representation.

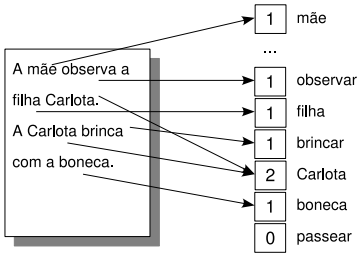

Figure 2. The original document and the corresponding bag-of-words representation.

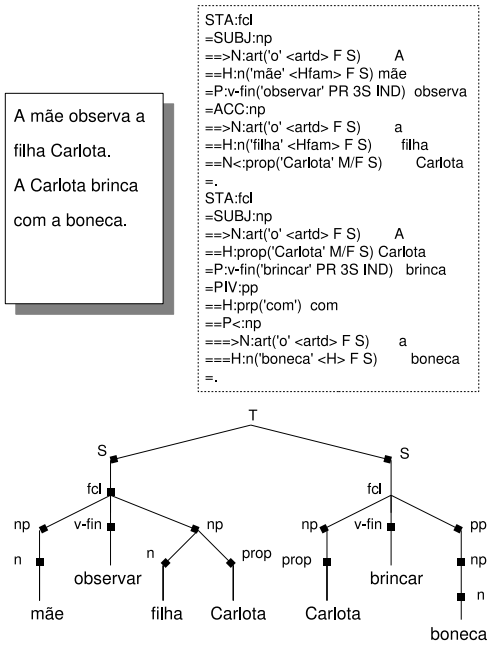

Figure 3. Original document, PALAVRAS output and syntactic-tree representation.

4. Experiments

This section introduces the used dataset, describes the experimental setup and presents the obtained results.

4.1. Dataset description

`Publico` corpus contains the Público newspaper daily news (from the years of 1994 an 1995) taken from 9 dif-

ferent sections (used as semantic classes). It totals 101646 documents, where there are 282657 distinct words, and, on average, 512 running words (tokens) and 254 unique words (types) per document.

For the syntactic experiments, a subset of `Publico` corpus with the October 1995 news was used. `Pub9510` has 4290 documents, with 70743 distinct words, and, on average, 215 tokens and 124 types per document. Table 1 shows the semantic classes and proportion of documents for each dataset.

	`Publico`	`Pub9510`
section	doc %	doc %
ciências, tecnologia e educação (science, technology, education)	6.2	6.7
cultura (culture)	15.5	14.5
desporto (sports)	9.9	10.3
diversos (diverse)	8.2	8.1
economia (economy)	13.3	10.5
local (local)	17.2	21.3
mundo (world)	9.4	9.3
nacional (national)	9.2	10.3
sociedade (society)	11.2	9.1

Table 1. `Publico` and `Pub9510` corpora: classes and proportion of documents.

4.2. Experimental setup

Bag-of-words representations used a linear kernel while the syntactic-tree one was run with the subset tree kernel. WEKA was run with default parameters (normalized training data and c=1, the trade-off between training error and margin) and SVM^{light}-TK with L=0.001 (decay factor) and c=10. A train-and-test procedure was applied with 33% of documents used for testing.

Learner performance was analyzed through precision (π), recall (ρ) and F_1 (f_1) measures [13] of each category (obtained from classification contingency table: prediction *vs.* manual classification). For each one, we calculated the micro- ($^{\mu}$) and macro-averages (^{M}) and made significance tests regarding a 95% confidence level.

4.3. Morphological information

First we considered morphological information with the traditional bag-of-words representation. It's the typical representation used in Information Retrieval techniques, it serves as baseline experiment and allows to verify lemmatization role in the classification process.

We made experiments using original words and their lemma (`lm`). Word selection was made using three classes of experiments: stop-word elimination[1] (`st`), filtering function (word frequency: `fr` and mutual information: `mi`) and threshold value (`t`). To weight the selected terms we used the three usual components: document (term frequency: `t`), collection (without component: `x`, and inverse term frequency: `f`) and normalization (co-sin: `c`). All these options can be graphically represented in a 3-dimensional space with normalization, selection and weighting axes. In turn, selection and weighting techniques can also be represented in other three-dimensional spaces (The marked point on Figure 4 corresponds to stop-word elimination, using lemmas as indexing terms, mutual information as filtering function, threshold value equals to one and `tfidf` as weighting technique).

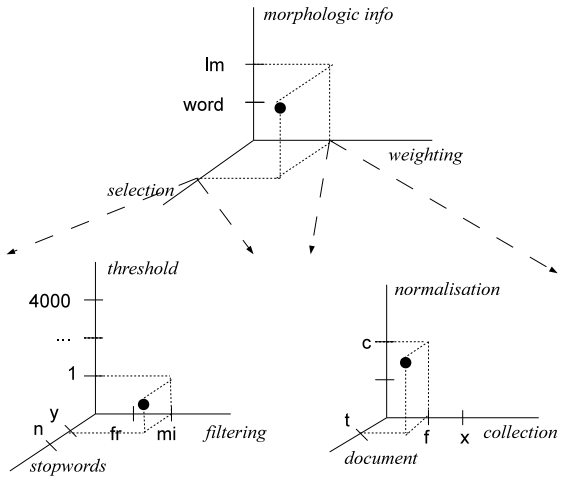

Figure 4. Graphical 3D representation of the experiments.

4.3.1 Results

We made experiments with combinations of the options described above (for `lemma` experiments we eliminated stopwords) with threshold values ranging from 1 to 4000 (this value indicates the smallest frequency above which the term is selected), in a total of 88 different runs. Table 2 shows the minimum, maximum, average and standard deviation values for the micro- and macro averages of the three performance measures.

To choose a representing experiment, we searched, for each performance measure, the ones with performance values having no significant difference with the maximum. There were 5 experiments with all 6 performance measures in the best set:

- word lemmatization (`lm`), with threshold value 1 (`t`$_1$);

[1] The stop-list was composed of 207 words.

	π^{μ}	ρ^{μ}	f_1^{μ}	π^{M}	ρ^{M}	f_1^{M}
min	.787	.787	.787	.779	.770	.774
max	.843	.843	.843	.842	.831	.836
avg	.824	.824	.824	.821	.810	.815
std	.012	.012	.012	.013	.013	.013

Table 2. `Publico` **minimum, maximum, average and standard deviation values for micro- and macro averages.**

- stop-word elimination (`st`), term frequency filtering function (`fr`) and threshold value 50 ($\texttt{t}_{50}$);
- stop-word elimination, mutual information filtering function (`im`), `tfidf` weighting technique (`tfc`) and threshold value 50 ($\texttt{t}_{50}$).

Table 3 shows their values (boldface values have no significant difference). From this set we chose $\texttt{lm-fr-tfc-t}_1$ as the representative experiment. It counts 205155 distinct words with averages of 288 tokens and 189 types per document.

	π^{μ}	ρ^{μ}	f_1^{μ}	π^{M}	ρ^{M}	f_1^{M}
$\texttt{lm-fr-txc-t}_1$	**.840**	**.840**	**.840**	**.839**	**.828**	**.833**
$\texttt{lm-fr-tfc-t}_1$	**.843**	**.843**	**.843**	**.842**	**.831**	**.836**
$\texttt{st-fr-txc-t}_{50}$	**.839**	**.839**	**.839**	**.837**	.826	**.831**
$\texttt{st-fr-tfc-t}_{50}$	**.840**	**.840**	**.840**	**.838**	**.828**	**.832**
$\texttt{st-im-tfc-t}_{50}$	**.840**	**.840**	**.840**	**.839**	**.828**	**.833**

Table 3. `Publico` **performance values for the morphological "best" experiments.**

4.4. Morphosyntactic information

To access the discriminative power of morphosyntactic information over classes we made a set of experiments including the word classes considered more informative – name (`n`), proper name (`prop`), adjective (`adj`) and verb (`v`).

4.4.1 Results

Using the same setting as the morphological information selected experiment we tried 15 different word class combinations. Table 4 shows the performance values for these experiments. Values of the morphological experiment, $\texttt{lm-fr-tfc-t}_1$, are also shown. From all experiments, only the one that combines all word classes (`n+prop+adj+v`) has equivalent values to the morphological one. This experiment counts 201327 distinct words with 250 tokens and 169 types per document.

We also made some experiments using higher thresholds, but performance values were lower.

	π^{μ}	ρ^{μ}	f_1^{μ}	π^{M}	ρ^{M}	f_1^{M}
$\texttt{lm-fr-tfc-t}_1$	**.843**	**.843**	**.843**	**.842**	**.831**	**.836**
`adj`	.683	.683	.683	.671	.663	.666
`n`	.808	.808	.808	.806	.704	.706
`prop`	.784	.784	.784	.778	.766	.771
`v`	.635	.635	.635	.623	.614	.618
`n+adj`	.807	.807	.807	.803	.792	.796
`n+prop`	.829	.829	.829	.824	.814	.818
`n+v`	.803	.803	.803	.821	.789	.802
`prop+adj`	.804	.804	.804	.800	.788	.793
`adj+v`	.732	.732	.732	.726	.715	.720
`prop+v`	.805	.805	.805	.802	.789	.795
`n+prop+adj`	.833	.833	.833	.830	.819	.824
`n+adj+v`	.813	.813	.813	.809	.800	.804
`n+prop+v`	.835	.835	.835	.833	.822	.827
`prop+adj+v`	.816	.816	.816	.811	.801	.805
`n+prop+adj+v`	**.839**	**.839**	**.839**	**.838**	**.827**	**.831**

Table 4. `Publico` **performance values for morphosyntactic experiments.**

4.5. Syntactic information

Even if our initial expectations about using syntactic information for text classification were low, we made some experiments with it.

Besides using the syntactic-tree representation (`tre`), and aiming to access the discriminating power of the structured representation, we considered other representations with information retrieved from the parse trees: a *bag-of-words* representation (`bag`) and a *sequence-of-words* representation (`seq`, an ordered tree, where words are root children).

4.5.1 Results

For this level of information we made two different experiments: one included the all document sentences (`tot`) and the other only the ones considered more informative – finite clauses with subject, predicate and direct object (`fcl`). For the `fcl` setting, we also made experiments including only the first n sentences of each document (trying to access if the first sentences have all the necessary information to classify news documents), with $n \in \{1, 3, 5, 10\}$. Table 5 shows the obtained performance measures.

The `bag.tot` experiment (in italics) has the best significant values for all measures. For the other experiments, we present in boldface the values with no significant difference when compared the second best value of each measure.

		π^{μ}	ρ^{μ}	f_1^{μ}	π^{M}	ρ^{M}	f_1^{M}
tre	tot	**.812**	**.812**	**.812**	**.810**	**.788**	**.792**
	fcl	.789	.789	.789	.789	.765	.770
	$\texttt{fcl}_1$	.675	.674	.674	.544	.526	.530
	$\texttt{fcl}_3$	.699	.699	.699	.683	.667	.670
	$\texttt{fcl}_5$	.734	.734	.734	.720	.702	.706
	$\texttt{fcl}_{10}$	.782	.782	.782	.776	.757	.760
seq	tot	**.829**	**.829**	**.829**	**.821**	**.811**	**.814**
	fcl	**.819**	**.819**	**.819**	**.814**	**.801**	**.804**
	$\texttt{fcl}_1$	.672	.671	.671	.549	.534	.539
	$\texttt{fcl}_3$	.708	.708	.708	.698	.681	.686
	$\texttt{fcl}_5$	.761	.761	.761	.750	.738	.741
	$\texttt{fcl}_{10}$	.791	.791	.791	.779	.771	.772
bag	tot	*.857*	*.857*	*.857*	*.858*	*.842*	*.847*
	fcl	**.824**	**.824**	**.824**	**.823**	**.809**	**.811**
	fcl	.824	.824	.824	.823	.809	.811
	$\texttt{fcl}_1$	.613	.613	.613	.605	.588	.594
	$\texttt{fcl}_3$	.734	.734	.734	.725	.707	.711
	$\texttt{fcl}_5$	.790	.790	.790	.782	.765	.768
	$\texttt{fcl}_{10}$	**.815**	**.815**	**.815**	**.803**	**.793**	**.793**

Table 5. `Pub9510` **performance measures for syntactic experiments.**

5. Evaluation

Looking at Table 3 one can say that it was possible to reduce the number of attributes ($\texttt{t}_{50}$) without compromising performance. However, these values were achieved only for experiments with the original words and not with lemmatization. It also seems that the mutual information filtering function should be used with `tfidf` weighting, while when filtering by the term frequency one, the weighting function is indifferent.

Concerning the morphosyntactic information study (Table 4), it was possible to achieve an equivalent performance to the "best" morphological setup when combining all four word classes together (`n+prop+adj+v`).

Taking into account the syntactic information experiments (Table 5) it is possible to say that structured representations introduce noise to text classification problems, since the best results were obtained using the bag-of-words representation. It also seems that adding information about sentence constituents and grammatical word class (in a structured way) damages the learner.

Since syntactic information experiments were made with the `Pub9510` corpus, another SVM was run with it using the morphological "best" setup ($\texttt{lm-fr-tfc-t}_1$). Table 6 reproduces the "best" syntactic experiment along with the obtained `Pub9510` morphological performance measures.

Values from the morphological setup are significantly better than the syntactic ones. In this way, and as expected, sentence syntactic structure does not properly reveal document class.

	π^{μ}	ρ^{μ}	f_1^{μ}	π^{M}	ρ^{M}	f_1^{M}
Morphological	.855	.855	.855	.854	.840	.844
Syntactic	.812	.812	.812	.810	.788	.792

Table 6. `Pub9510` **morphological and syntactic performance measures.**

6. Conclusions and Future Work

This paper presents a series of experiments, applied to the Portuguese language, aiming at verifying the value of incorporating linguistic information on text classification problems.

Concerning morphological information, results show that, when properly combined, word normalization, filtering and weighting functions and threshold values can sharpen performance. Comparing `Publico` dataset results with previous work on text classification with Portuguese written documents [6], one can conclude that the best combination depends on the dataset (or its domain).

For the Portuguese language, the use of morphosyntactic information as a term selection function generates classifiers with equal performance as the use of traditional Information Retrieval techniques. Moreover, by selecting words that belong to the name, proper name, adjective and verb word classes it's possible to reduce the number of terms without decaying performance.

Results' analysis show that, when using syntactic information, structured representations (syntactic-tree and sequence-of-words) harm the learner. Further more, as initially expected, sentence syntactic structure does not properly reveal document class; it could perhaps expose document writer or the kind of used language (like generic as the newspaper documents vs. specific areas of knowledge like medical or juridic ones).

Previous work made on the English language also studied the impact of linguistic processing on text classification. Moschitti and Basili [12] used linguistic tokens – nouns, verbs and adjectives, proper nouns and complex nominals and tokens augmented with their POS tag in context and concluded that SVM global performances were slightly penalized by the use of NLP-derived features. Scott and Matwin [14] also could not improve bag-of-words results while trying phrase-based and hypernym representations.

Regarding future work, we intend to perform further tests on different collections/domains and languages. It will be important to evaluate if these results are bound to the Portuguese language and/or the kind of the dataset domain.

Moreover, we intend to address the document representation problem by trying document representations that incorporate its semantic information.

References

[1] J. Backus. The syntax and semantics of the proposed international algebraic of the Zurich ACM-GAMM Conference. In *Proceedings of the International Conference on Information Processing – IFIP Congress*, pages 125–132. UNESCO, Paris, 1959.

[2] E. Bick. *The Parsing System PALAVRAS – Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework*. Aarhus University Press, 2000.

[3] E. Bick. A constraint grammar based question answering system for portuguese. In *Proceedings of the 11th Portuguese Conference of Artificial Intelligence – EPIA'03*, pages 414–418. LNAI Springer Verlag, 2003.

[4] N. Chomsky. Three models for the description of language. *IRI Transactions on Information Theory*, 2(3):113–124, 1956.

[5] M. Collins and N. Duffy. New ranking algorithms for parsing and tagging: Kernels over discrete structures, and the voted perceptron. In *ACL-02, 30th Annual Meeting of the Association for Computational Linguistics*, pages 263–270, 2002.

[6] T. Gonçalves, C. Silva, P. Quaresma, and R. Vieira. Analysing part-of-speech for Portuguese text classification. In *CICLing-06, 7th international Conference on Intelligent Text Processing and Computational Linguistics*, pages 551–562, Mexico City, MX, February 2006. Springer-Verlag.

[7] D. Haussler. Convolution kernels on discrete structures. Technical report, Department of Computer Science, University of California at Santa Cruz, 1999.

[8] T. Joachims. Making large-scale SVM learning practical. In B. Schölkopf, C. Burges, and A. Smola, editors, *Advances in Kernel Methods - Support Vector Learning*. MIT Press, 1999.

[9] J. Lopes, N. Marques, and V. Rocio. Polaris: POrtuguese Lexicon Acquisition and Retrieval Interactive System. In *The Practical Applications of Prolog*, page 665. Royal Society of Arts, 1994.

[10] A. Moschitti. A study on convolution kernels for shallow semantic parsing. In *ACL-04, 42nd Annual Meeting on Association for Computational Linguistics*, pages 335–342, Barcelona, SP, 2004.

[11] A. Moschitti. Efficient convolution kernels for dependency and constituent syntactic trees. In *ECML-06, 17th European Conference on Machine Learning*, pages 318–329, Berlin, DE, 2006.

[12] A. Moschitti and R. Basili. Complex linguistic features for text classification: a comprehensive study. In *ECIR-04, 26th European Conference on Information Retrieval Research*, pages 181–196, Sunderland, UK, 2004.

[13] G. Salton and M. McGill. *Introduction to Modern Information Retrieval*. McGraw-Hill, 1983.

[14] S. Scott and S. Matwin. Feature engineering for text classification. In *ICML-99, 16th International Conference on Machine Learning*, pages 379–388. Morgan Kaufmann Publishers, 1999.

[15] J. Shawe-Taylor and N. Cristianini. *Kernel Methods for Pattern Analysis*. Cambridge University Press, 2004.

[16] V. Vapnik. *The nature of statistical learning theory*. Springer, NY, 1995.

[17] V. Vapnik. *Statistical learning theory*. Wiley, NY, 1998.

[18] S. Vishwanathan and A. Smola. Fast kernels on strings and trees. In S. Becker, S. Thrun, and K. Obermayer, editors, *Advances in Neural Information Processing Systems*, volume 15, pages 569–576. MIT Press, Cambridge, MA, 2003.

[19] I. Witten and E. Frank. *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann, San Francisco, US, 2nd edition, 2005.

[20] D. Zhang and W. Lee. Question classification using support vector machines. In *SIGIR-03, 26th ACM International Conference on Research and Developement in Information Retrieval*, pages 26–32, 2003.