Text classification using tree kernels and linguistic information

Teresa Gonçalves, Paulo Quaresma
{tcg,pq}@di.uevora.pt

Dep. Informática, Universidade de Évora
7000-671 Évora, Portugal

Abstract

Standard Machine Learning approaches to text classification use the bag-of-words representation of documents to deceive the classification target function. Typical linguistic structures such as morphology, syntax and semantic are completely ignored in the learning process. This paper examines the role of these structures on the classifier construction applying the study to the Portuguese language.

Classifiers are built using the SVM algorithm on a newspaper's articles dataset. The results show that syntactic structure is not useful for text classification (as initially expected), but a novel structured representation that uses document's semantic information has the same discriminative power over classes as the traditional bag-of-words one.

1 Introduction

With the rapid growth of the World Wide Web, the task of classifying natural language documents in a set of predefined semantic classes became one of the key methods for organising on-line information. This task, usually called text classification, is a cornerstone in a wide area of applications. For example, Yahoo! directories classify Web pages by topic, on-line newspapers are tailored to the user reading preferences and automatic line routing agents address electronic messages to the appropriate expert.

A Machine Learning approach can be used to automatically build the classifiers. The construction process can be seen as a problem of supervised learning: the algorithm receives a relatively small set of labelled documents and generates the classifier. However, as learning algorithms do not directly interpret digital documents, it is required to transform each one in a compact representation of its contents. The most common approach, called bag-of-words, uses a statistical representation of the document, counting, in any way, its words. Language structures (such as syntax and semantic) typical of natural language documents are completely neglected.

To access the value of syntactic and semantic information we developed document representations that use parse trees and Discourse Representation Structures (DRS) from the Discourse Representation Theory [8] (DRT), respectively. The results obtained by these approaches were then compared with the ones obtained with common document representation techniques (that usually use some morphological information). To build the learners we used the Support Vector Machine (SVM) algorithm since it is known to produce good results on text classification tasks.

Document semantic information has also been used by [] to improve information retrieval in the web.

This paper is organised as follows: Section 2 points out the used techniques and datasets, Section 3 describes the document representation built for each studied level and Section 4 presents and evaluates the experiments. Conclusions and future work are pointed out in Sections 5 and 6.

2 Methods and Materials

This section describes linguistic information representations, the SVM algorithm and kernel functions for text classification, the used natural language tools to pre-process documents and the text classification software. It concludes by detailing the studied dataset and the experimental setup.

2.1 Linguistic information representation

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).

Most syntactic language representations are based on the context-free grammar (CFG) formalism introduced by [3] and, independently, by [1]: given a sentence, it generates the corresponding syntactic structure. Usually it's represented using a tree structure known as sentence's *parse tree* that contains its constituents structure (such as noun and verb phrases) and words' grammatical class.

On the other way, part of the semantic information can be obtained by context independent sentence meaning. This information is built by examining words' meaning and combining them. It can be produced directly from sentence's syntactic structure, and is named sentence's *logical form*.

Discourse Representation Theory [8] (DRT) is a dynamic semantic theory that uses a language over Discourse Representation Structures (DRS) to represent dependent context meaning. A simple DRS, is a pair of a set of Discourse Referents (DR) known as its universe, and a set of conditions. Intuitively the universe collects the discourse entities, while the conditions express entity restrictions (properties, relations). Figure 1 shows a DRS representation for the sentence "Joel Serrão ceases functions at Gulbenkian Institute of Science": there are three referents, x, y and z, and five conditions over them. x refers the name "Joel Serrão", y the function, and z the name "Gulbenkian Institute of Science" while the other conditions represent the action cease (being x the subject and y the object) and the relation at.

$x\ \ y\ \ z$
name(x,Joel_Serrão)
cease(x,y)
function(y)
rel(at,y,z)
name(z,Gulbenkian_Institute_of_Science)

Figure 1. DRS for the sentence "Joel Serrão ceases functions at Gulbenkian Institute of Science".

2.2 Support Vector Machine

Support Vector Machine (SVM) is an algorithm introduced by Vapnik [14] motivated by the theoretical results from the statistical learning theory. It joins a kernel technique with the structural risk minimisation 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 [13]. Its key aspects can be highlighted as follows (illustrated in Figure 2):

- 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.

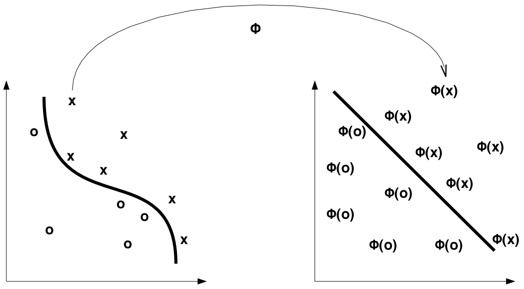


Figure 2. Kernel function: data's nonlinear pattern transformed into linear feature space.

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

2.3 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's frequency in each document) can be naturally reinterpreted as a kernel method. Normalisation and term reduction approaches can also be interpreted as kernel functions (see [13]) and standard kernels, like the polynomial one apply non linear transformations to the usual VSM approach.

The convolution kernel [6] 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 string, tree and graph structures. The convolution kernel definition is based on kernels defined over structure's components.

For tree structured objects, the feature space is indexed by subtrees and similarity is based on counting common subtrees. The subset tree kernel [4] is one of such kernels that uses ordered labelled trees and counts subsets of common subtrees between two trees.

2.4 Natural language processing tools

PALAVRAS [2] parser was used to obtain the parse tree for each sentence. It was developed in the context of the VISL project by the Institute of Language and Communication of the University of Southern Denmark.

SIN2SEM [11] application transforms PALAVRAS's parse tree into the corresponding logical form using DRS. These structures are represented by a two term Prolog predicate: the list of referents and a set of conditions over them. Although having mechanisms to consider sentence's meaning in the context where it is produced, SIN2SEM builds DRSs considering each sentence independently.

2.5 Text classification software

SVM algorithm was run using SVM^{light}-TK [9]. This software is an extension to SVM^{light} [7], that uses convolution kernels to represent tree structures. It implements two different tree kernels: the subtree kernel [15] and the subset tree kernel [4]. 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 nodes.These kernels have produced good results on parse tree ranking [4] and predicate argument classification [10, 16].

2.6 Dataset description

Público is a Portuguese daily newspaper and Publico9510 corpus contains its October 1995 news. They were taken from 9 different sections (used as semantic classes) totalling 4290 documents, where there are 70743 distinct words, and, on average, 215 running words (tokens) and 124 unique words (types) per document. Table 1 displays the semantic classes and the proportion of documents for each one.

section	doc %
ciências, tecnologia e educação (science, technology and education)	6.7
cultura (culture)	14.5
desporto (sports)	10.3
diversos (diverse)	8.1
economia (economy)	10.5
local (local)	21.3
mundo (world)	9.3
nacional (national)	10.3
sociedade (society)	9.1

Table 1. Semantic classes and documents' proportion for the Publico9510 corpus.

2.7 Experimental Setup

Learner's performance was analysed through precision (π), recall (ρ) and F_1 (f_1) measures [12] of each class (obtained from classification's contingency table: prediction *vs.* manual classification). For each measure, we calculated the micro- ($^{\mu}$) and macro-averages (^{M}) and made significance tests regarding a 95% confidence level.

SVM^{light}-TK was run with L=0.001 (decay factor) and c=10 (trade-off between training error and margin) using a train-and-test procedure with 33% of documents for testing.

3 Document representation

To use document's syntactic and semantic information in the process of building a text classifier, it's necessary to define a specific kernel or to adapt the representation to an existing one. Since it's possible to adapt both, the syntactic and semantic information, into a tree structure, we, next, describe those adaptations.

3.1 Syntactic-tree representation

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

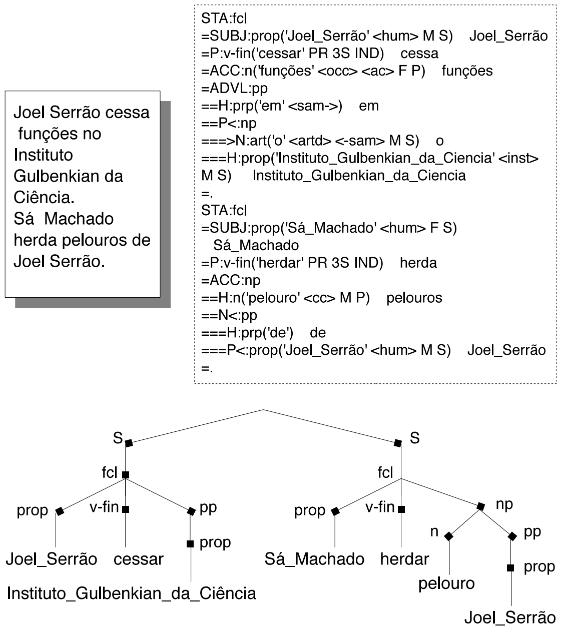
Nevertheless, we did not use the complete parse tree, but only the nodes of the following word classes: noun (n), proper noun (prop), adjective (adj), verb (v), pronoun (pron) and adverb (adv). Figure 3 illustrates a document with two sentences "Joel Serrão cessa funções no Instituto Gulbenkian da Ciência. Sá Machado herda pelouros de Joel Serrão." ("Joel Serrão ceases functions at Gulbenkian Institute of Science. Sá Machado inherits Manuel Serrão's portfolios."), PALAVRAS output and document's syntactic representation.

3.2 Discourse-structure representation

Being possible to represent a DRS by a tree structure, a document would be a tree with its children being all DRSs extracted from the document. We named this representation *discourse-structure* representation.

The tree structure that represents sentence's logical form can be obtained by applying substitutions on DRS referents confined by two kinds of specific conditions, one related with proper nouns and other with properties.

A proper name y associated with an entity x is represented by a condition name(x,y). By replacing all referent instances constrained by the name y (and removing the name condition) we obtain a corpus unification connecting all referents that refer to the same proper name y.



Joel Serrão cessa
funções no
Instituto
Gulbenkian da
Ciência.
Sá Machado
herda pelouros de
Joel Serrão.

```
STA:fcl
=SUBJ:prop('Joel_Serrão' <hum> M S)    Joel_Serrão
=P:v-fin('cessar' PR 3S IND)    cessa
=ACC:n('funções' <occ> <ac> F P)    funções
=ADVL:pp
==H:prp('em' <sam->)    em
==P<:np
===>N:art('o' <artd> <-sam> M S)    o
===H:prop('Instituto_Gulbenkian_da_Ciencia' <inst>
M S)     Instituto_Gulbenkian_da_Ciencia
=.
STA:fcl
=SUBJ:prop('Sá_Machado' <hum> F S)
   Sá_Machado
=P:v-fin('herdar' PR 3S IND)    herda
=ACC:np
==H:n('pelouro' <cc> M P)    pelouros
==N<:pp
===H:prp('de')    de
===P<:prop('Joel_Serrão' <hum> M S)    Joel_Serrão
=.
```

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

A property associated with an entity x is represented by an atomic condition $prop\,(x)$. By replacing all referent instances by $prop$ (and removing the $prop$ condition) we obtain a sentence unification. Since the same referent can be restricted by more than one property, we should build a list of referent properties and use it in the replacement.

Since the same referent x can be restricted by $\texttt{name}\,(x,\texttt{y})$ and $prop\,(x)$ conditions, it is always necessary to replace a referent by a list. For example, Figure 4 displays the DRS transformation when applying these two kinds of substitution to the same two sentences[1].

$x\ y\ z$		$a\ b\ c$

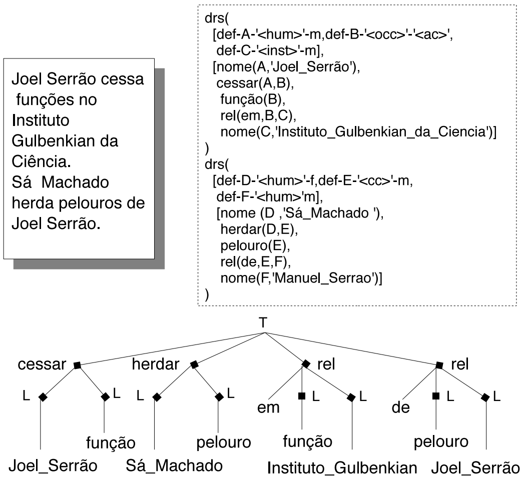
name(x,Joel_Serrão)
cease(x,y)
function(y)
rel(at,y,z)
name(z,Gulbenkian)
	+	name(a,Sá_Machado)
		inherit(a,b)
		portfolio(a)
		rel(of,b,c)
		name(c,Joel_Serrão)

↓

cease({Joel_Serrão},{function})
rel(at,{function},{Gulbenkian})
inherit({Sá_Machado},{portfolio})
rel(of,{portfolio},{Joel_Serrão})

Figure 4. DRS representation before and after applying referent replacement.

With these substitutions documents can be represented by a tree: children's root are document's DRS; DRS's conditions are, in turn, their children and below is the substitution list for each referent. Figure 5 illustrates the same two sentence document, SIN2SEM output and the semantic representation with both substitutions.



Joel Serrão cessa
funções no
Instituto
Gulbenkian da
Ciência.
Sá Machado
herda pelouros de
Joel Serrão.

```
drs(
  [def-A-'<hum>'-m,def-B-'<occ>'-'<ac>',
   def-C-'<inst>'-m],
  [nome(A,'Joel_Serrão'),
   cessar(A,B),
    função(B),
    rel(em,B,C),
    nome(C,'Instituto_Gulbenkian_da_Ciencia')]
)
drs(
  [def-D-'<hum>'-f,def-E-'<cc>'-m,
   def-F-'<hum>'m],
   [nome (D ,'Sá_Machado '),
    herdar(D,E),
    pelouro(E),
    rel(de,E,F),
    nome(F,'Manuel_Serrao')]
)
```

Figure 5. Original document, SIN2SEM output and discourse-structure representation.

4 Experiments

This section describes and presents the results obtained when using the syntactic and semantic information of the Publico95 dataset. Last subsection makes some evaluation and compares the results with the ones obtained on previous work using morphological information.

4.1 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 and aiming to access the discriminative power of the structured representation we considered other representations with information retrieved from the parse trees: a *bag-of-words* representation and a *sequence-of-words* representation (an ordered tree, where words are root's children).

4.1.1 Results

For this level of information we made two different experiments, one included the all document's sentences (tot) and other with 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 2 shows the obtained performance measures.

[1] Predicate names were translated for easier understanding.

		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
	fcl$_1$	.675	.674	.674	.544	.526	.530
	fcl$_3$	.699	.699	.699	.683	.667	.670
	fcl$_5$	.734	.734	.734	.720	.702	.706
	fcl$_{10}$	.782	.782	.782	.776	.757	.760
seq	tot	**.829**	**.829**	**.829**	**.821**	**.811**	**.814**
	fcl	**.819**	**.819**	**.819**	**.814**	**.801**	**.804**
	fcl$_1$	.672	.671	.671	.549	.534	.539
	fcl$_3$	.708	.708	.708	.698	.681	.686
	fcl$_5$	.761	.761	.761	.750	.738	.741
	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
	fcl$_1$	.613	.613	.613	.605	.588	.594
	fcl$_3$	.734	.734	.734	.725	.707	.711
	fcl$_5$	.790	.790	.790	.782	.765	.768
	fcl$_{10}$	**.815**	**.815**	**.815**	**.803**	**.793**	**.793**

Table 2. Performance measures for experiments using syntactic information.

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 with the second best value obtained for each measure.

4.2 Semantic Information

Besides using the discourse-structure representation and aiming to access its influence on the classification process we also considered a bag-of-words representation with the words extracted from the structured one.

4.2.1 Results

For the discourse-structure representation (dis) we considered two different kinds of referent substitutions; the ones connected with proper nouns (noun) and with proper nouns and property conditions (noun+pro). For each one we also tried to use the first n DRSs of each document with $n \in \{1, 3, 5, 10\}$. Table 3 shows the obtained performance measures. We present in boldface the values with no significant difference when compared with best value obtained for each measure.

4.3 Evaluation

Comparing the syntactic information experiments (Table 2) 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's constituents and grammatical word class (in a structured way) damages the learner.

		π^μ	ρ^μ	f_1^μ	π^M	ρ^M	f_1^M
noun	dis	.655	.655	.655	.732	.599	.623
	dis$_1$	.364	.364	.364	.533	.278	.288
	dis$_3$	.484	.484	.484	.604	.418	.451
	dis$_5$	.545	.545	.545	.660	.481	.510
	dis$_{10}$	.593	.593	.593	.692	.538	.567
	bag	**.821**	**.821**	**.821**	**.816**	**.808**	**.810**
noun +pro	dis	**.833**	**.833**	**.833**	**.831**	**.817**	**.820**
	dis$_1$	.471	.471	.471	.484	.437	.445
	dis$_3$	.679	.679	.679	.671	.645	.651
	dis$_5$	.740	.740	.740	.735	.710	.717
	dis$_{10}$	.787	.787	.787	.780	.772	.773
	bag	**.814**	**.814**	**.814**	**.822**	.788	.788

Table 3. Performance measures for experiments using semantic information.

On the other hand, using a structured representation for the semantic experiments (Table 3) with the proper noun and property substitutions (noun+pro) seems to add valuable information when compared to the bag-of-words representation (it has better macro- recall and f_1 values).

For comparing different linguistic levels of information we elected a "best" experiment. For the syntactic and semantic levels, we chose between those with a (full) structured representation being the syntactic-tree representation with all trees (tre.tot) and the discourse-structure representation with proper nouns and property conditions substitutions (dis.noun+pro).

To compare with traditional text classification approaches, we used the "best" experiment obtained on previous work [5]: it uses words' morphological information on a bag-of-words approach and uses word's lemma and TFIDF weighting measure with co-sin normalisation.

Table 4 displays the performance values of the "best" experiment for each linguistic information level. Boldface values points to values with no significant differences.

	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
Semantic	**.833**	**.833**	**.833**	**.831**	**.817**	**.820**

Table 4. Performance measures for each linguistic information level.

Since there is no significant difference between all the performance values concerning the morphological and semantic levels it is possible to say that both representations have the same discriminative power over classes. Moreover, its also possible to say that semantic information uses a valid form of attribute selection since it has 46186 types while the traditional bag-of-words approach has 70743.

5 Conclusions

This paper presents a series of experiments aiming at comparing our proposal of incorporating linguistic information using structured representations with the usual methods adopted on text classification problems.

Results analysis shows that, for the Portuguese language, when using syntactic information, structured representations (syntactic-tree and sequence-of-words) harm the learner. On the other hand, when using semantic information this is not the case: the discourse-structure representation with proper noun and property substitutions presents the same discriminative power as the non-structured one.

Moreover, while, as initially expected, the syntactic information shows lower results, the traditional bag-of-words approach (morphological information) and the proposed use of semantic information, show equivalent performance values. This statement demonstrates that both representations, one based on statistics over words and other based on document's meaning, are valid.

Considering the number of types used by morphological and semantic representations, one can conclude that documents' logical form performs a valid form of attribute selection: an about 30% reduction was accomplished.

Finally, one can conclude that the proposed discourse-structure representation is able to contain document's logical form and seems promissing since at this time it only describes document's meaning partially. We believe that by perfecting document's logical form, the semantic representation performance will be higher than the morphological one.

6 Future work

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.

On the other way, it is possible to obtain document's semantic representation closer to its real meaning by eliminating some of the known limitations of the used natural language tools. Although always generating an output, even in presence of incorrect or incomplete sentences, in some situations (for example, in presence of interrogative sentences) PALAVRAS generates incorrect parse trees. These errors are then propagated to DRS generation since parse trees become the input of SIN2SEM. SIN2SEM can also be refined by removing some of its limitations.

Finally, by incorporating other linguistic information like synonymous words, anaphora resolution and identification of named entities, DRS's set of conditions and referents would diminish, translating more accurately the true meaning of documents.

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

[4] 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.

[5] T. Gonçalves and P. Quaresma. Using linguistic information to classify portuguese text documents. In *MICAI-08*, 2008. submitted.

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

[7] 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.

[8] H. Kamp and U. Reyle. *From Discourse to Logic: An Introduction to Model Theoretic Semantics of Natural Language, Formal Logic and Discourse Representation Theory*, volume 42 of *Studies in Linguistics and Philosophy*. Springer, 1993.

[9] 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.

[10] 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.

[11] P. Quaresma, L. Quintano, I. Rodrigues, and P. Salgueiro. The University of Évora approach to QA@CLEF-2004. In *Multilingual Information Access for Text, Speech and Images*, Lecture Notes in Computer Science LNCS 3491, pages 534–543. Springer, 2005.

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

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

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

[15] 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.

[16] 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.