	QC		RTE	
λ	$DTK_{\boxtimes}$	$DTK_{\boxdot}$	$DTK_{\boxtimes}$	$DTK_{\boxdot}$
0.2	0.993	0.994	0.997	0.998
0.4	0.980	0.989	0.990	0.961
0.6	0.908	0.880	0.890	0.350
0.8	0.644	0.377	0.469	0.039
1.0	0.316	0.107	0.169	0.000

Table 2. Spearman's correlation between DTK values and TK values. Test trees were taken from the QC corpus in table (a) and the RTE corpus in table (b).

For QC, we used a standard question classification training and test set[2], where the test set are the 500 TREC 2001 test questions. To measure the task performance, we used a question multi-classifier by combining n binary SVMs according to the ONE-vs-ALL scheme, where the final output class is the one associated with the most probable prediction.

For RTE we considered the corpora ranging from the first challenge to the fifth (Dagan et al., 2006), except for the fourth, which has no training set. These sets are referred to as RTE1-5. The dev/test distribution for RTE1-3, and RTE5 is respectively 567/800, 800/800, 800/800, and 600/600 T-H pairs. We used these sets for the traditional task of pair-based entailment recognition, where a pair of text-hypothesis $p = (t, h)$ is assigned a positive or negative entailment class. For our comparative analysis, we use the syntax-based approach described in (Moschitti & Zanzotto, 2007) with two kernel function schemes: (1) $PK_S(p_1, p_2) = K_S(t_1, t_2) + K_S(h_1, h_2)$; and, (2) $PK_{S+Lex}(p_1, p_2) = Lex(t_1, h_1)Lex(t_2, h_2) + K_S(t_1, t_2) + K_S(h_1, h_2)$. Lex is a standard similarity feature between the text and the hypothesis and K_S is realized with TK, $DTK_{\boxtimes}$, and $DTK_{\boxdot}$. In the plots, the different PK_S kernels are referred to as TK, $DTK_{\boxtimes}$, and $DTK_{\boxdot}$ whereas the different PK_{S+Lex} kernels are referred to as $TK + Lex$, $DTK_{\boxtimes} + Lex$, and $DTK_{\boxdot} + Lex$.

5.2.2. Correlation between TK and DTK

As a first measure of the ability of DTK to emulate the classic TK, we considered the Spearman's correlation of their values computed on the parse trees for the sentences contained in QC and RTE corpora. Table 2 reports results and shows that DTK does not approximate adequately TK for $\lambda = 1$. This highlights the difficulty of DTKs to correctly handle pairs of large *active forests*, i.e., trees with many subtrees with weights around 1. The correlation improves dramatically when parameter λ is reduced. We can conclude that DTKs efficiently approximate TK for the $\lambda \leq 0.6$. These values are relevant for the applications as we will also see in the next section.

[2]The QC set is available at http://l2r.cs.uiuc.edu/~cogcomp/Data/QA/QC/

5.2.3. Task-based comparison

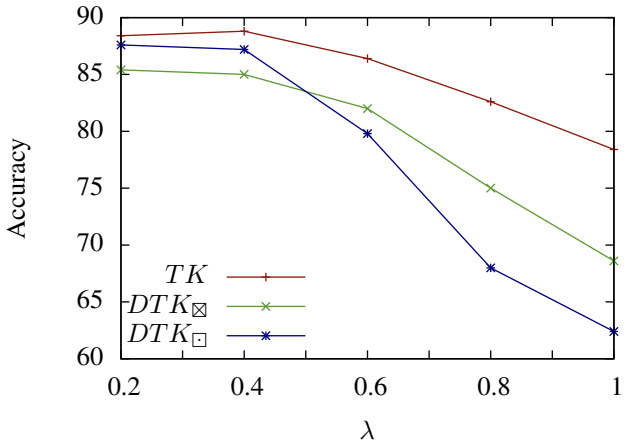

Figure 3. Performance on Question Classification task ($DTK_{\boxtimes}$ and $DTK_{\boxdot}$ rely on vectors of $d = 8192$).

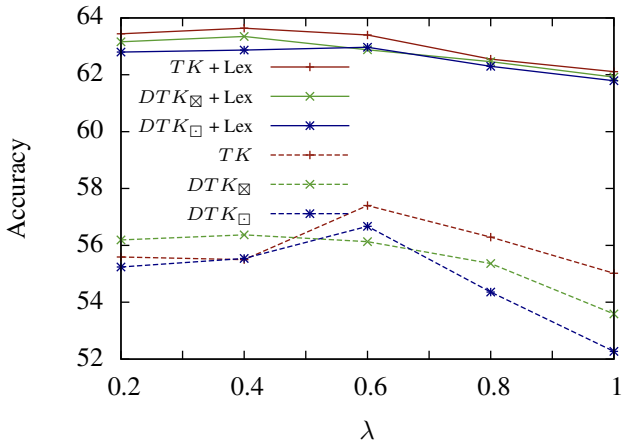

Figure 4. Performance on Recognizing Textual Entailment task ($DTK_{\boxtimes}$ and $DTK_{\boxdot}$ rely on vectors of $d = 8192$). Each point is the average of accuracy on the 4 data sets.

We performed both QC and RTE experiments for different values of parameter λ. Results are shown in Fig. 3 and 4 for QC and RTE tasks respectively.

For QC, DTK leads to worse performances with respect to TK, but the gap is narrower for small values of $\lambda \leq 0.4$ (with $DTK_{\boxdot}$ better than $DTK_{\boxtimes}$). These λ values produce better performance for the task. For RTE, for $\lambda \leq 0.4$, $DTK_{\boxtimes}$ and $DTK_{\boxdot}$ is similar to TK. Differences are not statistically significant except for for $\lambda = 0.4$ where $DTK_{\boxtimes}$ behaves better than TK (with $p < 0.1$). Statistical significance is computed using the two-sample Student t-test. $DTK_{\boxtimes} + Lex$ and $DTK_{\boxdot} + Lex$ are statisti-