Year	Dataset	Compared methods										Our method
		PROJ	PP-tfidf	DAN	RNN	LSTM	ST	GloVe	PSL	iRNN	SCBOW	ACVT
2012	MSRpar	0.44	0.41	0.40	0.19	0.09	0.15	0.48	0.43	0.43	0.44	**0.58**
	MSRvid	0.74	0.79	0.70	0.66	0.71	0.42	0.64	0.60	0.72	0.45	**0.83**
	SMTeuroparl	0.49	**0.53**	0.43	0.44	0.44	0.35	0.48	0.42	0.47	0.45	0.43
	OnWN	0.70	**0.73**	0.66	0.63	0.56	0.30	0.65	0.63	0.70	0.64	0.70
	SMTnews	0.63	**0.66**	0.60	0.51	0.51	0.31	0.50	0.57	0.58	0.39	0.54
2013	headlines	0.73	0.74	0.74	0.66	0.48	0.35	0.64	0.64	0.72	0.65	**0.77**
	OnWN	0.68	0.75	0.64	0.55	0.56	0.10	0.49	0.48	0.69	0.66	**0.85**
	FNWN	0.47	0.50	0.43	0.31	0.68	0.30	0.64	0.38	0.43	0.23	**0.50**
2014	deft-forum	0.54	**0.54**	0.49	0.42	0.46	0.13	0.23	0.33	0.49	0.41	0.48
	deft-news	0.72	0.74	0.72	0.56	0.68	0.24	0.68	0.61	0.72	0.68	**0.74**
	headlines	0.71	0.71	0.69	0.58	0.64	0.38	0.65	0.65	0.70	0.64	**0.72**
	images	0.78	0.81	0.77	0.68	0.63	0.51	0.64	0.63	0.78	0.63	**0.81**
	OnWN	0.86	0.81	0.79	0.68	0.62	0.23	0.68	0.61	0.78	0.61	**0.87**
	tweet-news	0.76	**0.77**	0.74	0.58	0.48	0.40	0.51	0.65	0.77	0.73	0.75
2015	answers-forums	0.65	0.68	0.63	0.34	0.54	0.36	0.31	0.39	0.63	0.22	**0.69**
	answers-students	0.78	0.79	0.78	0.65	0.66	0.34	0.62	0.69	0.78	0.37	**0.79**
	belief	0.75	**0.78**	0.72	0.52	0.63	0.25	0.41	0.59	0.78	0.48	0.70
	headlines	0.75	0.77	0.74	0.65	0.67	0.41	0.62	0.69	0.75	0.22	**0.79**
	images	0.80	**0.84**	0.78	0.71	0.64	0.18	0.68	0.70	0.81	0.26	0.82

Table 2: The comparison results; the bold number highlights one of strongest results in each dataset.

tains the forum post sentences, and the 15' belief dataset contains the Belief pairs for which their source documents are English Discussion Forum data. It is easy to understand that people usually write sentences in forums without using rigorous syntactic format, and so the grammars used in these sentences could be not guaranteed; and particularly sentences are also doped with a large number of colloquial terms and network abbreviations. These factors lead to the construction of syntactic structure inaccurate, weakening the performance of our model.

Classic methods: As same as to [Pilehvar and Navigli, 2015], we here also use the SemEval-2012 Semantic Similarity task to compare these classic methods. Table 3 lists the compared results. It can be seen from the table, our method beats all these methods for almost all these datasets. This further demonstrates the competitiveness of our model. Note that, as for the SMTeuroparl dataset, our model is significantly inferior than ADW (i.e., the performance gap is about 0.12). The reason is the same as our previous analysis. That is, this dataset contains much more *numerical items* and *special characters* for which our model lacks for the strong ability to model.

Impact of parameters: Recall Section 3.2, our model is involved with two important parameters μ and λ, where μ is a decay factor for the height of the tree, and λ is a decay factor for the length of the child sequences. We here study the impact of these two parameters on the accuracy of our model. Note that, in this set of experiments, we also test two other methods: one is known as CT which did not incorporate the word embedding technique and the attention weight mechanism; the other is known as CVT, which did not incorporate the attention weight mechanism. This way, it might be helpful for us to study the effectiveness of various techniques used in our model. To compute the similarity, CT uses PTK, while CVT uses the simple version of ACVT kernel in which the "weight" part is removed by setting "$Att_{weight} = 1$ in Equation 3". Next, we use the representative results to analyze the performance. Specifically, Figure 4 shows the compared results, these results are obtained by using the 13' OnWN dataset.

Dataset	Compared methods							Our method
	JCN	WUP	LCH	LIN	RES	ESA	ADW	ACVT
MSRvid	0.65	0.64	0.67	0.70	0.73	0.74	0.80	**0.83**
MSRpar	0.44	0.44	0.45	0.45	0.46	0.44	0.56	**0.58**
SMTeuroparl	0.20	0.21	0.18	0.23	0.25	0.48	**0.55**	0.43
OnWN	0.53	0.55	0.53	0.57	0.59	0.62	0.63	**0.70**
SMTnews	0.26	0.28	0.27	0.28	0.30	0.40	0.40	**0.54**

Table 3: The comparison between our method and classic methods.

It can be seen from Figure 4 that ACVT basically outperforms CVT, and CVT basically outperforms CT. This demonstrates that the word embedding technique and the attention weight mechanism are useful when we combine them together. As for ACVT, we can see from Figure 4(a) that, it has the best performance when we set $\mu = 0.2$ or 0.1 (compared to other settings such as $\mu = 0.9$). On the other hand, from Figure 4(b) we can see that our model can obtain best performance when we set $\lambda = 0.1$. These facts justify our default settings for parameter μ and λ, recall Section 4.1.

One could be curious why the curve of ACVT goes down when μ (resp., λ) increases. The main reason could be the followings. When the parameter μ (resp., λ) turns smaller, nodes near to the leaf level (resp., nodes with long child sequences) shall be penalized much more. This way, it makes our model pay more attention to the key information of sentences, which usually located at the upper layers of the ACV-tree. As such,

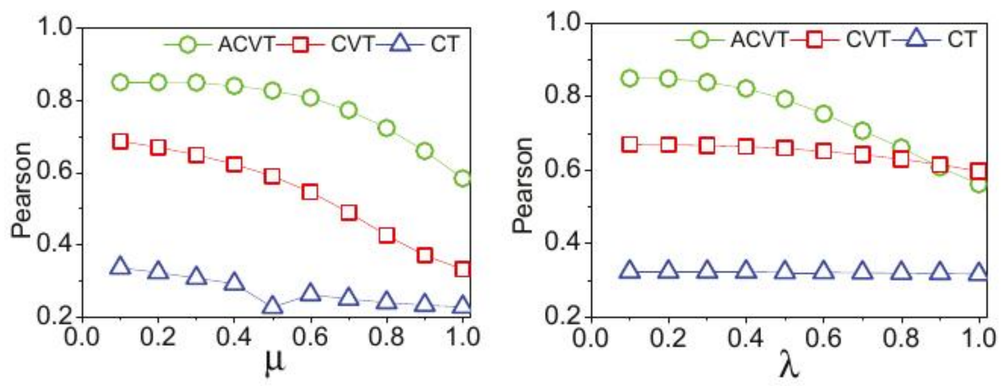

Figure 4: Varying μ and λ on the 13' OnWN dataset.