LATENT SEMANTIC RATIONAL KERNELS FOR TOPIC SPOTTING ON SPONTANEOUS CONVERSATIONAL SPEECH

Chao Weng, Biing-Hwang (Fred) Juang

Center for Signal and Image Processing, Georgia Institute of Technology, Atlanta, USA
{chao.weng, juang}@ece.gatech.edu

ABSTRACT

In this work, we propose latent semantic rational kernels (LSRK) for topic spotting on spontaneous conversational speech. Rather than mapping the input weighted finite-state transducers (WFSTs) onto a high dimensional n-gram feature space as in n-gram rational kernels, the proposed LSRK maps the WFSTs onto a latent semantic space. Moreover, with the LSRK framework, all available external knowledge can be flexibly incorporated to boost the topic spotting performance. The experiments we conducted on a spontaneous conversational task, Switchboard, show that our method can achieve significant performance gain over the baselines from 27.33% to 57.56% accuracy and almost double the classification accuracy over the n-gram rational kernels in all cases.

Index Terms— topic spotting, rational kernels, LSA, WFSTs

1. INTRODUCTION

Topic spotting aims at automatically determining the topics of the given speech utterances, which can be considered as a classification problem if the topics to be estimated are among a fixed set. Most of the previous works deal with this problem by first decoding the given speech utterances into transcripts and then treating it as a document categorization problem. Thus many text analysis techniques can be applied. In [1], a set of keywords are first selected according to their relative contribution to the discrimination for the topics and topic spotting is then employed by scoring the decoded transcript using those selected keywords. Similar idea has been applied to the famous AT&T HMIHY call-routing task [2], the concept of *salient words or phrases* was proposed [3] which are chosen with relative high mutual information with certain call-types, and then the calls are classified with the detection of those salient grammar fragments. More recently in [4], topic spotting with more sophisticated document classification algorithm, BOOSTEXTER, was explored; the authors also introduced a special learned grammar for the automatic speech recognition (ASR) decoding.

The common drawback of these methods is that the topic spotting strategy is still based on the 1-best ASR decoded transcript, which may not be reliable enough to deliver a good topic classification performance in some challenging tasks, e.g., spontaneous conversational speech. To overcome this, Cortes et al. [5] proposed the rational kernels, which are a series of kernels defined based on the weighted finite-state transducers (WFSTs). The topic classification can be conducted via support vector machine (SVM) with the rational kernels based on WFSTs (lattices) which compactly represent all the most likely transcripts from ASR outputs. Among all the rational kernels that have positive definite and symmetric (PDS) property, the n-gram rational kernel is prevalent in the topic spotting applications. The approach typically first maps the WFSTs to a high dimensional n-gram feature space and then employs an inner product for topic identification. However, the n-gram rational kernel assumes an exact match of the n-grams and treats contribution of each n-gram (words or phrases) to the topic discrimination uniformly resulting in substantial degradation in the topic spotting performance especially on some spontaneous speech in which filler or functional words frequently appear and interfere with the actual discriminability.

In this work, based on the n-gram rational kernels, we propose latent semantic rational kernels (LSRK) for topic spotting on spontaneous speech. Rather than mapping the WFSTs onto an n-gram feature space, we map the WFSTs onto a reduced dimensional latent semantic space as in latent semantic analysis (LSA) [6]. Under the WFSTs framework, compared to the n-gram rational kernels, LSRK needs another WFST's composition with the term-term similarity matrix and we generalize LSRK with respect to this similarity matrix such that any forms of external knowledge can be flexibly incorporated into the proposed LSRK framework to enhance the topic spotting performance. We will show that the n-gram rational kernels is a special case of LSRK when the term similarity matrix is an identity matrix. We conduct the topic spotting experiments using SVM with LSRK on a challenging task, the Switchboard, and it is shown that with LSRK we can achieve significant topic spotting performance gain over n-gram rational kernels from 27.33% to 57.56% classification accuracy. The remainder of this paper is organized as follows: Section 2 gives an overview of WFSTs and n-gram rational kernels, which serves as the preliminaries and background of this work. We will describe the formulations, detailed algorithms and the generalization of LSRK in Section 3. We report experimental results in Section 4 and finally conclude our work by making a brief discussion on how the paper's contributions are related to prior work in Section 5.

2. N-GRAM RATIONAL KERNELS

In this section, we will present some WFSTs algebraic definitions and notations needed to introduce rational kernels and describe the n-gram rational kernel.

2.1. WFSTs and Rational Kernels

A system $(\mathbb{K}, \oplus, \otimes, \overline{0}, \overline{1})$ is a semiring if : $(\mathbb{K}, \oplus, \overline{0})$ is a commutative monoid with identity element $\overline{0}$; $(\mathbb{K}, \otimes, \overline{1})$ is a monoid with identity element $\overline{1}$; $\otimes$ distributes over $\oplus$; and $\overline{0}$ is an annihilator for $\otimes$ (for all $a \in \mathbb{K}$, $a \otimes \overline{0} = \overline{0} \otimes a = 0$). We list some commonly used semirings in Table 1. Two semirings that are often used in the speech and language processing applications are the log semirings (similar to the probability semiring but with weight manipulation conducted in the negative log domain) and the tropical semirings (derived from the log semiring used for approximate

SEMIRING	SET	$\oplus$	$\otimes$	$\overline{0}$	$\overline{1}$
Boolean	$\{0,1\}$	$\vee$	$\wedge$	0	1
Probability	$\mathbb{R}_+$	$+$	$\times$	0	1
Log	$\mathbb{R} \cup \{-\infty, +\infty\}$	$\oplus_{\log}$	$+$	$+\infty$	0
Tropical	$\mathbb{R} \cup \{-\infty, +\infty\}$	min	$+$	$+\infty$	0

Table 1. Commonly used Semirings. $\oplus_{\log}$ is defined by $x \oplus_{\log} y = -\log(e^{-x} + e^{-y})$.

Viterbi decoding). A WFST T [7] over a semiring $\mathbb{K}$ is an 8-tuple $T = (\Sigma, \Delta, Q, I, F, E, \lambda, \rho)$, where Σ is the finite input alphabet of the transducer, Δ is the finite output alphabet, Q is a finite set of states, $I \subseteq Q$ is the set of initial states, $F \subseteq Q$ is the set of final states, $E \subseteq Q \times (\Sigma \cup \{\epsilon\}) \times (\Delta \cup \{\epsilon\}) \times \mathbb{K} \times Q$ is a finite set of transitions, $\lambda : I \to \mathbb{K}$ is the initial weight function, and $\rho : F \to \mathbb{K}$ is the final weight function mapping F to $\mathbb{K}$. A weighted finite-state acceptor (WFSA) can be formally defined in a similar way but with the same input and output labels. Given a transition $e \in E$, we denote by $p[e]$ its origin or previous state and $n[e]$ its destination or next state, and $w[e]$ its weight. A path $\pi = e_1 \cdots e_k$ consists of consecutive transitions, $n[e_{i-1}] = p[e_i]$, $i = 2, ..., k$, and a successful path in a WFST/WFSA is a path from an initial state to a final state with the weight as the $\otimes$-product of the weights of its constituent transitions, $w[\pi] = w[e_1] \otimes \cdots \otimes w[e_k]$. Let $P(q, q')$ be the set of paths from state q to q' and $P(q, x, y, q')$ the set of paths from q to q' with input label $x \in \Sigma$ and output label $y \in \Delta$, then the output weight associated by T to any pair of input-output string (x, y) is given by,

$$[\![T]\!](x, y) = \bigoplus_{\pi \in P(I, x, y, F)} \lambda(p[\pi]) \otimes w[\pi] \otimes \rho[n[\pi]], \quad (1)$$

which is well defined in $\mathbb{K}$ and $[\![T]\!](x, y) = \overline{0}$ when $P(I, x, y, F) = \emptyset$. Given a weighted automaton or transducer M, the *shortest-distance* from state q to the set of final states F is defined as the $\oplus$-sum of all the paths from q to F,

$$d[q] = \bigoplus_{\pi \in P(q, F)} w[\pi] \otimes \rho[n[\pi]]. \quad (2)$$

For any transducer T, we denote by T^{-1} its *inverse*, that is the transducer by swapping the input and output labels of each transition and the input and output alphabets. For *composition*, let $T_1 = (\Sigma, \Delta, Q_1, I_1, F_1, E_1, \lambda_1, \rho_1)$ and $T_2 = (\Delta, \Omega, Q_2, I_2, F_2, E_2, \lambda_2, \rho_2)$ be two WFSTs defined over a commutative semiring $\mathbb{K}$ such that Δ, the output alphabet of T_1, coincides with the input alphabet of T_2. Then, the result of the composition of T_1 and T_2 is a weighted transducer $T_1 \circ T_2$ and for all input-output strings pair (x, y),

$$[\![T_1 \circ T_2]\!](x, y) = \bigoplus_{z \in \Delta} [\![T_1]\!](x, z) \otimes [\![T_2]\!](z, y). \quad (3)$$

Note that a transducer can be viewed as a matrix over the set $\Sigma \times \Delta$ and composition as the corresponding matrix-multiplication.

Let A be a weighted automaton defined over the semiring $\mathbb{K}$ and the alphabet Σ, B a weighted automaton defined over the semiring $\mathbb{K}$ and the alphabet Δ, a weighted transducer $T = (\Sigma, \Delta, Q, I, F, E, \lambda, \rho)$ over the semiring $\mathbb{K}$ and a function $\psi : \mathbb{K} \to \mathbb{R}$. Then the *rational kernels* $K(A, B)$ over A and B is given by,

$$K(A, B) = \psi \left(\bigoplus_{(x,y) \in \Sigma \times \Delta} [\![A]\!](x) \otimes [\![T]\!](x, y) \otimes [\![B]\!](y) \right), \quad (4)$$

for convenience, we use $w[M]$ as the shorthand for the *shortest distance* from the start state I to the set of final states F of the transducer M, Eq.(4) thus can be written as,

$$K(A, B) = \psi \left(\bigoplus_{(x,y) \in \Sigma \times \Delta} [\![A \circ T \circ B]\!](x, y) \right) = \psi(w[A \circ T \circ B]) \quad (5)$$

2.2. N-gram Rational Kernels

An N-gram kernel is a rational kernel that has PDS property and has been successfully and widely used in speech or text classification applications [8]. Suppose A is a WFST (word lattice) output from an ASR system, which evaluates a probability distribution P_A over all strings that can be represented by A, $s \in \Sigma$. Modulo a normalization constant, the weight assigned by A to a string x is $[\![A]\!](x) = -\log P_A(x)$ (for the log semiring). Denote by $|s|_x$ the number of occurrences of a sequence x in the string s. The expected count or number of occurrences of an *n-gram* sequence x in s for the probability distribution P_A is,

$$c(A, x) = \sum_s P_A(s)|s|_x. \quad (6)$$

The n-gram rational kernel k_n for two WFSTs A_1 and A_2 is defined as,

$$k_n(A_1, A_2) = \sum_{|x|=n} c(A_1, x)c(A_2, x), \quad (7)$$

which is typically the sum of product of the expected counts that A_1 and A_2 assign to their common *n-gram* sequences. In the WFST framework, n-gram rational kernels can be calculated efficiently as,

$$k_n(A_1, A_2) = w[(A_1 \circ T) \circ (T^{-1} \circ A_2)] = w[A_1 \circ (T \circ T^{-1}) \circ A_2], \quad (8)$$

where T is the transducer that can be used to extract all n-grams and compute $c(A_1, x)$,

$$T = (\Sigma \times \{\epsilon\})^* (\sum_{x \in \Sigma} \{x\} \times \{x\})^n (\Sigma \times \{\epsilon\})^*. \quad (9)$$

Fig.1 shows the T transducer in the case of bi-gram sequences ($n = 2$) and for the vocabulary $\Sigma = \{a, b\}$.

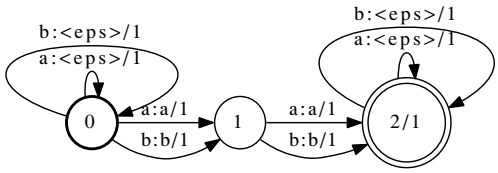

Fig. 1. T transducer computing expected counts of bi-gram sequences of a word lattice with $\Sigma = \{a, b\}$, note that $\langle$eps$\rangle$ represents ϵ denoting the empty label

3. LATENT SEMANTIC RATIONAL KERNELS

In this section, based on the n-gram rational kernels, we propose a latent semantic rational kernel (LSRK) and show how LSRK can be generalized to incorporate any form of external knowledge to enhance the topic spotting performance.

3.1. Latent Semantic Rational Kernels Formulations

Recall that kernel methods first map the inputs to a high dimensional feature ϕ space, and take the inner product of them. Here we rewrite Eq.(7) as,

$$k_n(A_1, A_2) = \sum_{|x|=n} c(A_1, x)c(A_2, x) = \langle \phi(A_1), \phi(A_1) \rangle$$
$$= \phi(A_1)^T \phi(A_1), \quad (10)$$

where $\phi(A)$ is the mapped feature vector, we can see what n-gram rational kernels do is first mapping a WFST to an n-gram space, and the value corresponding to each dimension is the expected count for this n-gram. It can be seen that there are two main limitations with n-gram rational kernels used for topic spotting: The N-gram kernel assumes that WFSTs from the same topic share many exactly matched n-grams, but in reality many n-grams are often correlated, sometimes synonymous. Furthermore, the produced WFST assumes uniform contribution from the n-grams, while we often observe many words that are not useful for topic discrimination, e.g., filler or functional words, at the same time, some significant terms as salient phrases in HMIHY that represent certain topic well may have the risk of being neglected in the evaluation process.

If we treat WFST as *a distribution over multiple documents*, the ideas of both LSA and latent semantic kernels (LSK) [9] can be applied here naturally. In LSA, a document is first represented by a vertical vector d indexed by the terms in the vocabulary, and the corpus is then represented by a term-document matrix D, whose columns are indexed by the documents and whose rows are indexed by the terms, $D = [d_1, ..., d_m]$. If we define the kernels over two documents as,

$$K(d_1, d_2) = \langle d_1, d_2 \rangle = d_1^T d_2, \quad (11)$$

this is similar to n-grams rational kernels over WFST, which measures the similarity by counting exact matches terms/n-grams. But as in LSA or LSK, d will be first mapped into a latent semantic space to explore the semantic relationship between terms. This space with a much lower dimensionality is given by employing singular value decomposition (SVD) on the D matrix. Denote by $\mathcal{T}$ the linear transform we use to map d to the latent semantic space, the latent semantic kernel is defined as,

$$K(d_1, d_2) = \langle \mathcal{T}d_1, \mathcal{T}d_2 \rangle = d_1^T \mathcal{T}^T \mathcal{T} d_2. \quad (12)$$

Similarly, for the n-gram rational kernels, we can modify Eq.(10) to,

$$k_n(A_1, A_2) = \langle \mathcal{T}\phi(A_1), \mathcal{T}\phi(A_1) \rangle = \phi(A_1)^T \mathcal{T}^T \mathcal{T} \phi(A_1), \quad (13)$$

since we do not need to express the feature vector explicitly (kernel trick), we define the *Latent Semantic Rational Kernels* (LSRK) as,

$$k_n(A_1, A_2) = \langle \mathcal{T}\phi(A_1), \mathcal{T}\phi(A_1) \rangle = \phi(A_1)^T \mathcal{S} \phi(A_1), \quad (14)$$

compared with basic n-gram rational kernels, we only need to multiply the feature vector by one matrix $\mathcal{S}$ before employing the inner product, which implies another WFST composition operation. In the WFST framework, suppose S is the WFST representing the matrix $\mathcal{S}$, the LSRK can be calculated as,

$$k_n(A_1, A_2) = w[(A_1 \circ T) \circ S \circ (T^{-1} \circ A_2)]$$
$$= w[A_1 \circ (T \circ S \circ T^{-1}) \circ A_2] \quad (15)$$

where S WFST can be defined as,

$$S = (\{\epsilon\} \times \{\epsilon\})^* (\sum_{x \in \Sigma} \{x\} \times \{x\})^n (\{\epsilon\} \times \{\epsilon\})^*, \quad (16)$$

One example of the S transducer in the bi-gram case is shown in Fig.2, in which each arc corresponds to the elements in the $\mathcal{S}$ matrix; e.g., $\mathcal{S}(i, j)$ corresponds to the arc with input label i, output label j and weight $\mathcal{S}(i, j)$. Then, the S for n-gram LSRK is constructed by concatenating n stages like this. S may appear to contain a large number of arcs, $n \times |\Sigma| \times |\Sigma|$, but in reality $\mathcal{S}$ can be very sparse over the non-diagonal elements and is thus still tractable after we use some heuristics to prune it.

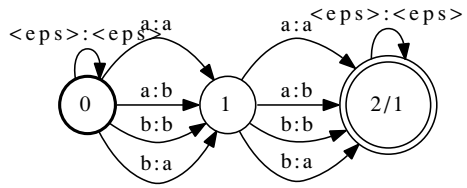

Fig. 2. S transducer (without weight on arcs) computing LSRK of a word lattice with $\Sigma = \{a, b\}$ (bi-gram case)

3.2. Generalization of Latent Semantic Rational Kernels

If we take an insightful look at the $\mathcal{S}$ matrix as in Eq.(14), it actually can be viewed as the *term-term similarity matrix* which specifies the *semantic similarity* between terms, e.g., the value of element $\mathcal{S}(i, j)$ measures the semantic similarity between term i and j. In the n-gram rational kernels case, it assumes semantic similarity of the same term is 1 and there exists no semantic similarity between different terms which corresponds to the special case of LSRK with $\mathcal{S}$ being set to identity matrix I. This motivates us to generalize the LSRK with respect to the term-term similarity matrix $\mathcal{S}$, i.e., $\mathcal{S}$ is not necessarily constructed from the LSA, instead it can be designed in multiple ways such that any form of available external knowledge can be incorporated into it. This generalization gives us lots of possibilities to use LSRK. We list several typical cases to use LSRK as illustrations.

- If $\mathcal{S} = I$, LSRK is equivalent to the n-gram rational kernels.
- If $\mathcal{S} = \mathrm{diag}\left(\mathrm{idf}^2(1), ..., \mathrm{idf}^2(i), ..., \mathrm{idf}^2(N)\right)$, where $\mathrm{idf}(i)$ is the *inverse document frequency* of term i according to the training corpus. In this case, LSRK will count the expected tf-idfs (term frequency-inverse document frequency) assigned to the common n-grams. Note that the expected term frequency is already evaluated with $A \circ T$ part in Eq.(15), we only need idfs for each term in the matrix $\mathcal{S}$.
- If $\mathcal{S} = U_K \Sigma_K^{-1} \Sigma_K^{-1} U_K^T$, where U_K and Σ_K are the corresponding matrices obtained from the K-rank approximation to the term-document matrix using SVD as in LSA, $D \approx U_K \Sigma_K V_K^T$. In this case, the $\mathcal{S}$ is constructed from the latent semantic space in a data-driven way.
- If $\mathcal{S}_{ij} = \mathrm{WordNet} :: \mathrm{Similarity}(i, j)$, the $\mathcal{S}$ matrix is then constructed from the WordNet ontology [10]. Various algorithms [11] using WordNet can be used to determine the similarity; this approach models the similarity based on the distance between the conceptual categories of words and the hierarchical structure in the WordNet.

In real applications, several techniques can be combined to obtain an effective $\mathcal{S}$ matrix. The training corpus we use to estimate the matrix $\mathcal{S}$ is not limited to the speech transcripts which usually are limited and expensive. With LSRK, more available text corpus can be utilized to boost the topic spotting performance.

4. EXPERIMENTS

We evaluated the proposed LSRK for topic spotting on a challenging conversational telephone speech task, Switchboard-1 Release 2, which is a collection of 2438 two-sided telephone conversations among 543 speakers (302 males, 241 females). Each pair of callers is introduced a topic for discussion and there are about 70 topics.

4.1. The ASR system and WFSTs (lattices) Generation

We first describe the ASR system we use to generate the WFSTs (lattices) for each utterances. The acoustic models are cross-word triphone models represented by 3-state left-to-right HMMs (5-state HMMs for silence) trained using MLE on about half data of the whole Switchboard corpus. A tri-gram language model (LM) is trained for decoding. The input features are MFCCs coupled with their linear discriminant analysis (LDA) and maximum likelihood linear transform (MLLT) and feature-space maximum likelihood linear regression (fMLLR) for speaker adaptation during later iterations. The WER of the ASR system on the HUB5 English evaluation set is 33.4%. With this ASR system, we first trained a uni-gram LM using the whole transcripts of the dataset and then use it to generate lattices for around 100K utterances (about half of the whole dataset). These 100K WFSTs are the data we would use for the following topic spotting experiment.

4.2. Topic Spotting with LSRK on the subset of Switchboard

It is found that a substantial amount of ill-formed utterances for topic spotting exist among those 100K utterances, e.g., "UH, YEAH". We first filter out the filler words, functional words and stop words from the transcripts for each utterance and then select utterances whose filtered transcripts have appropriate length. (We set the length threshold to 20, and there are around only 10K utterances left.) From those selected utterances, we filter out those topics that have less than 200 utterances. Finally, 4405 utterances on 19 topics are selected for the topic spotting tasks, and for each topic we randomly choose 90% for training and 10% for testing, as shown in Table 2.

We conduct the topic spotting on this subset of Switchboard using multiclass SVMs with n-gram rational kernels (baselines) and LSRK respectively. For the n-gram rational kernels, we conduct the experiments with different n. Note that when $n > 1$, the kernels are actually obtained by taking the sum of all k_m in Eq.(7) as $K_n = \sum_{m=1}^{n} k_m$, $1 \leq m \leq n$. For the LSRK, the way we generated $\mathcal{S}$ is a combination of LSA and tf-idf. We use each conversation transcript with those test utterances excluded as one document (2438 in total) to form the term-document matrix D, then use the tf-idf weights to scale the corresponding term in the matrix. Since the $\mathcal{S}$ is very large (over 30K$\times$30K), we pruned those non-diagonal elements by selecting most N significant elements. Note that $\mathcal{S}$ is symmetrical, so we can just focus on the upper-right half of the matrix, and choose most $N/2$ elements. With the pruned $\mathcal{S}$, we compile it into transducers S to employ the LSRK. As shown in Table.3, we get 27.33% and 28.22% classification accuracy (which are comparable to the numbers reported on the Switchboard in [4]) in the unigram and bigram cases (we omit the results for higher n because the further improvements are marginal). For the LSRK, we report the results in terms of different rank K for the LSA and the number of left non-diagonal elements N after pruning. As can be seen, in all cases we obtain significant topic spotting gain (almost doubled) over n-gram rational kernels baseline. And with less pruned $\mathcal{S}$, we can get higher accuracy and the highest one with 57.56%.

TOPIC	TRAIN	TEST	TOTAL
RECIPES/FOOD/COOKING	242	28	270
CAPITAL PUNISHMENT	197	23	220
PUBLIC EDUCATION	196	22	218
BUYING A CAR	207	24	231
PETS	204	23	227
WOMEN'S ROLE	191	22	213
TV PROGRAM	197	22	219
DIRECTIONS	245	28	273
GARDENING	200	23	223
WEATHER CLIMATE	250	28	278
MOVIES	193	22	215
GUN CONTROL	212	24	236
DRUG TESTING	193	22	215
AUTO REPAIRS	197	22	219
HOBBIES AND CRAFTS	188	21	209
EXERCISE AND FITNESS	230	26	256
AIR POLLUTION	180	21	201
CAMPING	186	21	207
RECYCLING	247	28	275
TOTAL	3955	450	4405

Table 2. Number of utterances (train/test/total) for each topic in the subset of Switchboard used for the topic spotting evaluation

System/Method	N (pruning)	K (LSA)	Accuracy
Unigram RK	-	-	27.33%
Bigram RK	-	-	28.22%
LSRK	40K$\times$2	500	52.44%
LSRK	80K$\times$2	500	52.89%
LSRK	120K$\times$2	500	52.44%
LSRK	160K$\times$2	500	54.00%
LSRK	200K$\times$2	500	53.78%
LSRK	1000K$\times$2	500	56.67%
LSRK	40K$\times$2	750	52.67%
LSRK	80K$\times$2	750	52.44%
LSRK	120K$\times$2	750	52.89%
LSRK	160K$\times$2	750	53.56%
LSRK	200K$\times$2	750	53.33%
LSRK	1000K$\times$2	750	**57.56%**

Table 3. Classification accuracies on the subset of Switchboard, N is the number of non-diagonal elements left in $\mathcal{S}$ after pruning, K is the rank for the low dimensional term-document matrix approximation in LSA.

5. CONCLUSIONS

We conclude this work by briefly discussing how the paper's contributions are related to prior work. To overcome the main drawback of the previous works [1][2][3][4] on topic spotting that the spotting is still based on the 1-best ASR decoded transcript, Cortes et al. [5] proposed the rational kernels and successfully applied one of rational kernels, n-gram rational kernels to this application. In this work, we proposed latent semantic rational kernels (LSRK) for topic spotting, rather than mapping WFSTs into n-gram high-dimension feature space, the proposed LSRK mapping WFSTs into a latent semantic space. Moreover, with the LSRK framework, all available external knowledge can be flexibly incorporated to boost the topic spotting performance. The experiments we conducted on a spontaneous conversational task, Switchboard, show that our method can achieve significant performance gain over the baselines, obtained almost the doubled classification accuracy over the n-gram rational kernels in all cases.

6. REFERENCES

[1] J. H. Wright, M. J. Carey, and E. S. Parris, "Improved topic spotting through statistical modelling of keyword dependencies," in *Proc. ICASSP1995*, 1995, pp. 313–316.

[2] A. L. Gorin, G. Riccardi, and J. H. Wright, "How may i help you?," *Speech Communication*, vol. 23, pp. 113–127, 1997.

[3] A. L. Gorin, G. Riccardi, and J. H. Wright, "Automatic acquisition of sailent grammar fragments for call-type classification," in *Proc. EuroSpeech97*, 1997.

[4] K. Myers, M. Kearns, S. Singh, and M. A. Walker, "A boosting approach to topic spotting on subdialogues," in *Proc. ICML00*, 2000, pp. 662–669.

[5] C. Cortes, P. Haffner, and M. Mohri, "Rational kernels: Theory and algorithms," *Journal of Machine Learning Research*, vol. 5, pp. 1035–1062, 2004.

[6] S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman, "Indexing by latent semantic analysis," *Journal of the American society for information science*, vol. 41, no. 6, pp. 391–407, 1990.

[7] M. Mohri, F. Pereira, and M. Riley, "Weighted finite-state transducers in speech recognition," *Computer Speech and Language*, vol. 20, no. 1, pp. 69–88, 2002.

[8] C. Cortes, P. Haffner, and M. Mohri, "Lattice kernels for spoken dialog classification," in *Proc. ICASSP03*, 2003, pp. 628–631.

[9] N. Cristianini, J. Shawe-Taylor, and H. Lodhi, "Latent semantic kernels," *Journal of Intellignet Information Systems*, vol. 18, no. 2-3, pp. 127–152, 2002.

[10] G. A. Miller, R. Beckwith, C. Fellbaum, D. Gross, and K. Miller, "Wordnet: An on-line lexical database," *International Journal of Lexicography*, vol. 3, pp. 235–244, 1990.

[11] T. Pedersen, S. Patwardhan, and J. Michelizzi, "Wordnet::similarity: measuring the relatedness of concepts," in *Demonstration Papers at HLT-NAACL 2004*, 2004, HLT-NAACL–Demonstrations '04, pp. 38–41.