# SupMMD: A Sentence Importance Model for Extractive Summarization using Maximum Mean Discrepancy

**Umanga Bista**[⋆] **Alexander Patrick Mathews**[⋆] **Aditya Krishna Menon**[‡] **Lexing Xie**[⋆]
[⋆]Australian National University, Canberra, ACT, Australia
[‡]Google Research, New York, NY, United States
[⋆]{umanga.bista,alex.mathews,lexing.xie}@anu.edu.au,
[‡]adityakmenon@google.com

## Abstract

Most work on multi-document summarization has focused on *generic* summarization of information present in each individual document set. However, the under-explored setting of *update summarization*, where the goal is to identify the *new* information present in each set, is of equal practical interest (e.g., presenting readers with updates on an evolving news topic). In this work, we present SupMMD, a novel technique for generic and update summarization based on the *maximum mean discrepancy* from kernel two-sample testing. SupMMD combines both supervised learning for salience and unsupervised learning for coverage and diversity. Further, we adapt multiple kernel learning to make use of similarity across multiple information sources (e.g., text features and knowledge based concepts). We show the efficacy of SupMMD in both generic and update summarization tasks by meeting or exceeding the current state-of-the-art on the DUC-2004 and TAC-2009 datasets.

## 1 Introduction

Multi-document summarization is the problem of producing condensed digests of salient information from multiple sources, such as articles. Concretely, suppose we are given two sets of articles (denoted set A and set B) on a related topic (e.g., climate change, the COVID-19 pandemic), separated by publication timestamp or geographic region. We may then identify three possible instantiations of multi-document summarization (see Figure 1):

(i) *generic summarization*, where the goal is to summarize a set (A, or B) individually.

(ii) *comparative summarization*, where the goal is to summarize a set (B) against another set (A) while highlighting the differences.

(iii) *update summarization*, where the goal is both generic summarization of set A *and* comparative summarization of set B versus A.

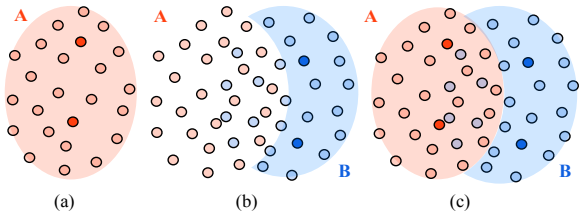


Figure 1: Different summarization tasks: (a) Generic (b) Comparative (c) Update. Two sets of articles (set A and B) are denoted by red and blue circles, respectively. Summary prototypes are bold circles, and information coverage of each tasks is filled by respective colors.

Most existing work on this topic has focused on the generic summarization task. However, update summarization is of equal practical interest. Intuitively, the comparative aspect of this setting aims to inform a user of new information on a topic they are already familiar with.

Multi-document extractive summarization methods can be unsupervised or supervised. *Unsupervised* methods typically define salience (or coverage) using a global model of sentence-sentence similarity. Methods based on retrieval (Goldstein et al., 1999), centroids (Radev et al., 2004), graph centrality (Erkan and Radev, 2004), or utility maximization (Lin and Bilmes, 2010, 2011; Gillick and Favre, 2009) have been well explored. However, sentence salience also depends on *surface features* (e.g., position, length, presence of cue words); effectively capturing these requires supervised models specific to the dataset and task. A body of work has incorporated such information through *supervised learning*, for example based on point processes (Kulesza and Taskar, 2012), learning important words (Hong and Nenkova, 2014), graph neural networks (Yasunaga et al., 2017), and support vector regression (Varma et al., 2009). These supervised methods have either a separate model for learning and inference, leading to a disconnect between learning sentence salience and sentence

selection (Varma et al., 2009; Yasunaga et al., 2017; Hong and Nenkova, 2014), or are designed specifically for generic summarization (Kulesza and Taskar, 2012). In this work, we propose *SupMMD*, which has a single model of learning salience and inference and can be applied to generic and comparative summarization. We make the following contributions:

**(1)** We present *SupMMD*, a novel technique for both generic and update summarization that combines supervised learning for salience and unsupervised learning for coverage and diversity. *SupMMD* has a single model for learning and inference.

**(2)** We adapt multiple kernel learning (Cortes et al., 2010) into our model, which allows similarity across multiple information sources (e.g., text features and knowledge based concepts) to be used.

**(3)** We show that *SupMMD* meets or exceeds the state-of-the-art in generic and update summarization on the DUC-2004 and TAC-2009 datasets.

## 2 Literature Review

Multi-document summarization can be *extractive*, where salient pieces of the original text such as sentences are selected to form the summary, or *abstractive*, where a new text is generated by paraphrasing important information. The former is popular as it often creates semantically and grammatically correct summaries (Nallapati et al., 2017). In this work, we focus on *generic* and *update multi-document* summarization in the *extractive* setting.

Most extractive summarizers have two components: sentence scoring and selection. A variety of unsupervised and supervised methods have been developed for the former. *Unsupervised* sentence scorers are based on centroids (Radev et al., 2004), graph centrality (Erkan and Radev, 2004), retrieval relevance (Goldstein et al., 1999), word statistics (Nenkova and Vanderwende, 2005), topic models (Haghighi and Vanderwende, 2009), or concept coverage (Gillick and Favre, 2009; Lin and Bilmes, 2011). *Supervised* techniques include: using a graph based neural network (Yasunaga et al., 2017), learning sentence quality from point processes (Kulesza and Taskar, 2012), combining word importances (Hong and Nenkova, 2014), combining sentence and phrase importances (Cao et al., 2015), or employing a mixture of submodular functions (Lin and Bilmes, 2012).

Sentence selection methods can be broadly categorized as *greedy* methods (Goldstein et al., 1999; Radev et al., 2004; Erkan and Radev, 2004; Nenkova and Vanderwende, 2005; Cao et al., 2015; Haghighi and Vanderwende, 2009; Hong and Nenkova, 2014; Kulesza and Taskar, 2012; Cao et al., 2015; Varma et al., 2009), which produce approximate solutions by iteratively selecting the sentences with the maximal score, or *exact* integer linear programming (ILP) based methods (Gillick and Favre, 2009; Cao et al., 2015). Some greedy methods use an objective which belongs to a special class of set functions called *submodular functions* (Lin and Bilmes, 2010, 2012, 2011; Kulesza and Taskar, 2012), which have good approximation guarantees under greedy optimization (Nemhauser et al., 1978).

There has been limited research into update and comparative summarization. Notable prior work includes maximizing concept coverage using ILP (Gillick et al., 2009), learning sentence scores using a support vector regressor (Varma et al., 2009), and temporal content filtering (Zhang et al., 2009). Bista et al. (2019) cast the comparative summarization problem as classification, and use MMD (Gretton et al., 2012). In this work, we adapt their method to learn *sentence importances* driven by surface features.

## 3 Summarization as Classification

We review a perspective introduced by Bista et al. (2019), where summarization is viewed as classification, and provide a brief introduction to Maximum Mean Discrepancy (MMD). Both these ideas form the basis of our subsequent method.

### 3.1 Generic Summarization as Classification

Let $\{V^{\mathrm{t}}\}_{\mathrm{t}=1}^{T}$ be $T$ topics of articles that we wish to summarize. For a topic $t$, we wish to select *summary sentences* $S^{\mathrm{t}}$. Bista et al. (2019) formulated summarization as selecting prototypes that minimize the accuracy of a powerful classifier between sentences in the input and summary. The intuition is that a powerful classifier should not be able to distinguish between the sentences from articles and summary sentences. Formally, we pick

$$S^{\mathrm{t}} = \underset{S \in \mathcal{S}^t}{\operatorname{argmax}} \; -\operatorname{Acc}(V^{\mathrm{t}}, S), \qquad (3.1)$$

where $\mathcal{S}^{\mathrm{t}} \subset 2^{V^{\mathrm{t}}} : \forall_{S' \in \mathcal{S}} \sum_{s \in S'} \operatorname{len}(s) \leq L$ comprise subsets of $V^{\mathrm{t}}$ with upto $L$ words, and $\operatorname{Acc}(X, Y)$ is the accuracy of the best possible classifier that distinguishes between elements in sets $X$ and $Y$; we shall shortly realize this using MMD.

### 3.2 Comparative Summarization as Competing Binary Classification

For comparative summarization between two sets $A$ and $B$, Bista et al. (2019) introduced an additional term into (3.1), giving rise to *competing goals* for the classifier: it should not be able to distinguish between the summaries and sentences from set $B$, but *should* be able to distinguish between the summaries and sentences from set $A$. Formally, let $V_B^{\mathrm{t}}$ be the set of sentences in set $B$, $V_A^{\mathrm{t}}$ be the sentences in set to compare (set $A$). Then, for suitable $\lambda > 0$, we seek $S^{\mathrm{t}}$, the summary sentences of set $B$,

$$S^{\mathrm{t}} = \underset{S \in \mathbb{S}^{\mathrm{t}}}{\operatorname{argmax}} \left[ -\mathrm{Acc}(V_B^{\mathrm{t}}, S) + \lambda \cdot \mathrm{Acc}(V_A^{\mathrm{t}}, S) \right]. \quad (3.2)$$

The hyperparameter $\lambda$ controls the relative importance of accurately representing articles in set B, versus not representing the articles in set A.

### 3.3 Maximum Mean Discrepancy (MMD)

The *MMD* is a kernel-based measure of the distance between two distributions. More formally:

**Definition 3.1.** Let $\mathcal{H}$ be a Reproducing Kernel Hilbert Space (RKHS) with associated kernel $k$. Let $\mathcal{F}$ be the set of functions $h : \mathcal{X} \mapsto \mathbb{R}$ in the unit ball of $\mathcal{H}$, where $\mathcal{X}$ is a topological space. Then, the MMD between distributions $p, q$ is the *maximal difference* in expectations of functions from $\mathcal{F}$ under $p, q$ (Gretton et al., 2012):

$$\mathrm{MMD}_{\mathcal{F}}(p, q) = \sup_{h \in \mathcal{F}} \left( \mathop{\mathbb{E}}_{x \sim p} [h(x)] - \mathop{\mathbb{E}}_{y \sim q} [h(y)] \right). \quad (3.3)$$

A small MMD value indicates that $p, q$ are similar. Given finite samples $X \sim p^n$ and $Y \sim q^m$, an empirical estimate of the MMD, denoted as $\mathrm{MMD}^2_{\mathcal{F}}(X, Y)$, can be computed as:

$$\frac{1}{n^2} \sum_{x,x'} k(x, x') + \frac{1}{m^2} \sum_{y,y'} k(y, y') - \frac{2}{n \cdot m} \sum_{x,y} k(x, y). \quad (3.4)$$

### 3.4 MMD for Summarization

The MMD corresponds to the minimal achievable loss of a centroi-based kernel classifier (Sriperumbudur et al., 2009). Consequently, we use $\mathrm{MMD}^2_{\mathcal{F}}(V, S)$ to approximate the $\mathrm{Acc}(V, S)$ in (3.1) and (3.2), using a suitable kernel $k$ that measures the similarity of two sentences. Intuitively, this selects summaries $S$ which best represent the *distribution* of original sentences $V$.

Note that if we expand $\mathrm{MMD}^2_{\mathcal{F}}(V, S)$ as per (3.4) and later in §4.6, the first term is irrelevant for optimization. The second and third term capture the coverage and diversity of the summary sentences without any supervision. Hence, this is an *unsupervised* summarization.

## 4 The SupMMD method

We start by developing a technique for incorporating sentence importance into MMD for the purpose of generic multi-document extractive summarization. We then extend this method to comparative summarization, and incorporate multiple different kernels to use a diverse sets of features.

### 4.1 From MMD to Weighted MMD

Unsupervised MMD (Bista et al., 2019) selects representative sentences that cover relevant concepts while retaining diversity. The notion of representativeness is based on a global model of sentence-sentence similarity; however, this notion of representativeness is not necessarily well matched to the selection of salient information. Salience of a sentence may be determined by *surface features* such as position in the article, or number of words. For example, news articles are often written such that sentences at the start of a article have the characteristics of a summary (Kedzie et al., 2018). Learning a notion of salience that is specific to the summarization task and dataset requires supervised training. Thus, we extend the MMD model by incorporating supervised *sentence importance weighting*.

Let $v, s \in \mathcal{X}$ be independent samples drawn from the distributions of article sentences $p$ and summary sentences $q$ on the space of all sentences $\mathcal{X}$. We define non-negative *importance functions* $f^p_\theta$, $f^q_\theta$ parameterized by learnable parameters $\theta$. We restrict these functions so that $\mathbb{E}_p f^p_\theta(v) = 1$ and $\mathbb{E}_q f^q_\theta(s) = 1$. Equipped with $f_\theta$, we may modify MMD such that the importance of sentences which are good summary candidates is increased.

**Definition 4.1.** The *weighted MMD* $\mathrm{MMD}_{\mathcal{F}}(p, q, \theta)$ between $p, q$ is

$$\sup_{h \in \mathcal{F}} \left( \mathop{\mathbb{E}}_{p} \left[ f^p_\theta(v) \cdot h(v) \right] - \mathop{\mathbb{E}}_{q} \left[ f^q_\theta(s) \cdot h(s) \right] \right) \quad (4.1)$$

Note that classic MMD (3.3) is a special case of (4.1) where $f_\theta \equiv 1$.

In practice, the supremum over all $h$ is impossible to compute directly. We thus derive an alternative form for Equation 4.1.

**Lemma 4.1.** For $\|h\|_{\mathcal{H}} \leq 1$, (4.1) is equivalently

$$\left\| \mathbb{E}_p[f^p_\theta(v) \cdot \phi(v)] - \mathbb{E}_q[f^q_\theta(s) \cdot \phi(s)] \right\|_{\mathcal{H}}. \quad (4.2)$$

In the above, $\phi : \mathcal{X} \mapsto \mathcal{F}$ is a canonical feature mapping of sentences and summaries from $\mathcal{X}$ to RKHS. The derivation, which mirrors a similar derivation for MMD (Gretton et al., 2012), is given in the Appendix.

### 4.2 Importance Function

We use log-linear models as importance functions, as they are a common choice of sentence importance (Kulesza and Taskar, 2012) and easy to fit when training data is scarce. Formally, the log-linear importance function is: $f_\theta(v) = \exp(\langle \theta, \omega(v) \rangle)$, where $\omega(v)$ is the surface features of sentence $v$. We can define the empirical estimates $f_\theta^{n_t}(v)$, $f_\theta^{m_t}(s)$ of the importance functions $f_\theta^p(v)$ and $f_\theta^q(s)$ as:

$$f_\theta^{n_t}(v) = \frac{f_\theta(v)}{\sum_{v' \in V^t} f_\theta(v')} \cdot n_t$$
$$f_\theta^{m_t}(s) = \frac{f_\theta(s)}{\sum_{s' \in S^t} f_\theta(s')} \cdot m_t \quad (4.3)$$

where $n_t = |V^t|$ is the number of sentences and $m_t = |S^t|$ is the number of summary sentences in topic t.

### 4.3 Training: Generic Summarization

The parameters $\theta$ of the log-linear importance function must be learned from data, so we define a loss function based on weighted MMD. Let $\{(V^t, S^t)\}_{t=1}^T$ be the $T$ training tuples. Then, the loss of topic $t$ is the square of importance weighted empirical MMD between sentences and summary sentences from within the topic:

$$\mathcal{L}^t(V^t, S^t, \theta) = \mathrm{MMD}^2_{\mathcal{F}}(V^t, S^t, \theta) \quad (4.4)$$

where $\mathrm{MMD}^2_{\mathcal{F}}(V^t, S^t, \theta)$ is an empirical estimate of the weighted $\mathrm{MMD}^2_{\mathcal{F}}(p, q, \theta)$. Applying the kernel trick to Equation 4.4 gives (see Appendix):

$$\mathcal{L}^t = \frac{1}{n_t^2} \sum_{v,v'} f_\theta^{n_t}(v) \cdot f_\theta^{n_t}(v') \cdot k(v, v') - \frac{2}{n_t \cdot m_t} \sum_{v,s} f_\theta^{n_t}(v) \cdot f_\theta^{m_t}(s) \cdot k(v, s) + \frac{1}{m_t^2} \sum_{s,s'} f_\theta^{m_t}(s) \cdot f_\theta^{m_t}(s') \cdot k(s, s') \quad (4.5)$$

Equation 4.5 is the loss for a single topic but during training we will instead minimize the average loss over all topics in the training set, i.e., $\min_\theta \frac{1}{T} \sum_{t=1}^T \mathcal{L}^t(V^t, S^t, \theta)$. Intuitively, we learn the parameters $\theta$ by minimizing an importance weighted distance between sentences and ground truth summary sentences over all topics.

### 4.4 Training: Comparative Summarization

We now extend the learning task to comparative summarization using the competing binary classifiers idea of Bista et al. (2019) (cf. §3.2). Specifically, we replace the accuracy terms in Equation 3.2 with the square of weighted MMD. Given the $T$ comparative training tuples $\{(V_B^t, V_A^t, S^t)\}_{t=1}^T$, then the objective is to minimize:

$$\min_{\theta_B, \theta_A} \frac{1}{T} \sum_t \left( \mathcal{L}^t(V_B^t, S^t, \theta_B) - \lambda \cdot \mathcal{L}^t(V_A^t, S^t, \theta_A) \right) \quad (4.6)$$

Note there are two sets of importance parameters $\theta_B, \theta_A$ one for each of the two document sets.

### 4.5 Multiple Kernel Learning

We employ Multiple Kernel Learning (MKL) to make use of data from multiple sources in our MMD summarization framework. We adapt two stage kernel learning (Cortes et al., 2010), where different kernels are linearly combined to maximize the alignment with the *target kernel* of the classification problem. Since MMD can be interpreted as classifiability (Sriperumbudur et al., 2009) MKL fits neatly into our MMD based summarization objective. Intuitively, MKL should identify a good combination of kernels for building a classifier that separates summary and non-summary sentences.

Let $\{k_i\}_{i=1}^p$ be $p$ kernel functions. For topic $t$, let $\mathbf{K}_i^t$ be the kernel matrix according to kernel function $k_i$, and $\overline{\mathbf{K}}_i^t = \mathbf{U}_{n_t} \mathbf{K}_i^t \mathbf{U}_{n_t}$ be the centered kernel matrix, with $\mathbf{U}_{n_t} = \mathbf{I} - \mathbf{1}\mathbf{1}^{\mathrm{T}}/n_t$. Let $\mathbf{y}^t = \{\pm 1\}^{n^t}$ be the ground truth summary labels with $y_i^t = +1$ iff $i \in S^t$. The *target kernel* $\mathbf{y}^t(\mathbf{y}^t)^T$ represents the ideal notion of similarity between sentences. The non-negative kernel weights $\mathbf{w}$ which lead to the optimal alignment with the target kernel are given by (Cortes et al., 2010)

$$\min_{\mathbf{w} \geq 0} \mathbf{w}^{\mathrm{T}} (\mathbf{M}^t)^{\mathrm{T}} \mathbf{w} - 2\mathbf{w}^{\mathrm{T}} \mathbf{a}^t, \quad (4.7)$$

where $\mathbf{M}^t \in \mathbb{R}^{p \times p}$ has $\mathbf{M}_{rs}^t = \langle \overline{\mathbf{K}}_r, \overline{\mathbf{K}}_s \rangle_{\mathrm{F}}$ and $\mathbf{a}^t \in \mathbb{R}^p$ has $a_i = \langle \overline{\mathbf{K}}_i^t, \mathbf{y}^t(\mathbf{y}^t)^{\mathrm{T}} \rangle_{\mathrm{F}}$.

The kernel function must be characteristic for MMD to be a valid metric (Muandet et al., 2017). Most popular kernels used for bag of words like text features (including TF-IDF), the linear kernel ($k(\mathbf{x}, \mathbf{y}) = \langle \mathbf{x}, \mathbf{y} \rangle$) and the cosine kernel

($k(\mathbf{x},\mathbf{y}) = \frac{\langle \mathbf{x},\mathbf{y} \rangle}{\|\mathbf{x}\|\|\mathbf{y}\|}$), are not characteristic (Sriperumbudur et al., 2010). Fortunately, the exponential kernel, $k(\mathbf{x},\mathbf{y}) = \exp(\gamma k'(\mathbf{x},\mathbf{y}))$, $\gamma > 0$, is characteristic for any kernel $k'$ (Steinwart, 2001). Hence, we use the *normalized exponential kernel* combined with the cosine kernel, $k(\mathbf{x},\mathbf{y}) = \exp(-\gamma)\exp\big(\gamma \sum_{i=1}^{p} w_i \cdot \cos\big(\mathbf{x}^{(i)},\mathbf{y}^{(i)}\big)\big)$.

### 4.6 Inference

Given a learned importance function $f_\theta$, we may find the best set of summary sentences $\bar{S}^t$ for generic summarization via:

$$\bar{S}^t = \underset{S \subset \mathcal{S}^t}{\operatorname{argmax}} \; -\mathcal{L}^t(V^t, S^t, \theta) \tag{4.8}$$

Similarly, for the comparative task, with learne importance functions, we seek $\bar{S}^t$ as:

$$\underset{S \subset \mathcal{S}^t}{\operatorname{argmax}} \left(-\mathcal{L}^t(V_B^t, S^t, \theta_B) + \lambda \mathcal{L}^t(V_A^t, S^t, \theta_A)\right) \tag{4.9}$$

Both these inference problems are budgeted maximization problems, which are often solved by greedy algorithms (Lin and Bilmes, 2010). The generic unsupervised summarization task is submodular and monotone under certain conditions (Kim et al., 2016), so greedy algorithms have good theoretical guarantees (Nemhauser et al., 1978). While our supervised variants do not have these guarantees, we find that greedy optimization nonetheless leads to good solutions.

## 5 Experimental setup

We include guidance on applying *SupMMD*, and the details required to reproduce our experiments.

### 5.1 Datasets

We use four standard multi-document summarization benchmark datasets: DUC-2003, DUC-2004, TAC-2008 and TAC-2009[1]; dataset statistics are provided in Table 1. Each of these datasets has multiple topics, where each topic in turn has multiple news articles and four human written summaries. In one setting we use DUC-2003 as the training set and DUC-2004 as test set, and in another setting we use TAC-2008 as the training set and TAC-2009 as the test set – both settings are common in the literature. The DUC datasets can be used for generic summarization while TAC, being an update summarization task, can be used for both generic (set A) and comparative summarization (set B).

[1] https://duc.nist.gov/data.html

### 5.2 Data Preprocessing and Preparation

The DUC and TAC datasets are provided as collections of XML documents, so it is necessary to extract relevant text and then perform sentence and word tokenization. For DUC we clean the text using various regular expressions the details of which are provided in our code release. We train `PunktSentenceTokenizer` to detect sentence boundaries, and use the standard NLTK (Bird, 2006) word tokenizer. For the TAC dataset, we use the preprocessing pipeline employed by Gillick et al. (2009) [2]. This enables a cleaner comparison with the state-of-the-art ICSI (Gillick et al., 2009) method on the TAC dataset. For all datasets, we keep the sentences between 8 and 55 words per Yasunaga et al. (2017).

### 5.3 Feature Representations

Our method requires two different sets of sentence features: *text features*, which are used to compute the sentence-sentence similarity as part of the kernel; and *surface features*, which are used in learning the sentence importance model.

#### 5.3.1 Text Features

Each sentence has three different feature representations: unigrams, bigrams and entities. The unigrams are stemmed words, with stop words from the NLTK english list removed. The bigrams are a combination of stemmed unigrams and bigrams. The entities are DBPedia concepts extracted using DBPedia spotlight (Mendes et al., 2011).

We use a Term Frequency Inverse Sentence Frequency (TF-ISF) (Neto et al., 2000) representation for all text features. TF-ISF has been used extensively in multi-document summarization (Dias et al., 2007; Alguliev et al., 2011; Wan et al., 2007).

#### 5.3.2 Surface Features

We use 10 surface features for the DUC dataset, and 12 for the TAC dataset:

**position**: There are five position features. Four indicators denote the 1st, 2nd, 3rd or a later position of the sentence in the article. The final feature gives the position relative to the length of the article.

**counts**: There are two count features: the number of words and number of nouns. We use the `spaCy` [3] part of speech tagging to find nouns.

**tfisf**: This is the sum of the TS-ISF scores for unigrams composing the sentence. For sentence

[2] https://github.com/benob/icsisumm

[3] https://spacy.io

| Dataset | # topics | # sents | Oracle (ours) avg summ sents | R1 | R2 | Oracle (Liu and Lapata, 2019) avg summ sents | R1 | R2 |
|---|---|---|---|---|---|---|---|---|
| DUC2003 | 30 | 6989 | 3.73 | 43.1 | 17.0 | 3.40 | 42.2 | 16.2 |
| DUC2004 | 50 | 12148 | 4.02 | 42.0 | 14.9 | 3.46 | 40.6 | 14.2 |
| TAC2008-A | 48 | 9914 | 3.90 | 45.5 | 19.4 | 3.42 | 44.0 | 18.6 |
| TAC2008-B | 48 | 9147 | 3.83 | 44.9 | 19.5 | 3.50 | 43.6 | 18.7 |
| TAC2009-A | 44 | 9509 | 4.07 | 46.9 | 20.5 | 3.32 | 44.5 | 19.1 |
| TAC2009-B | 44 | 8543 | 3.61 | 44.8 | 19.2 | 3.27 | 43.1 | 18.1 |

Table 1: Dataset statistics and oracle performance. We report the number of topics in each dataset, along with the number of sentences after preprocessing. We show the ROUGE scores of our oracle method and the one by Liu and Lapata (2019) with average number of sentence in summary from each method.

$s$, this is $\sum_{w \in s} \text{isf}(w) \cdot \text{tf}(w, s)$, where $\text{isf}(w)$ is the inverse sentence frequency of unigram $w$, and $\text{tf}(w, s)$ is the term frequency of $w$ in $s$.

**btfisf**: The boosted sum of TS-ISF scores for unigrams composing the sentence. Specifically, we compute $\sum_{w \in s} \text{isf}(w) \cdot b(w) \cdot \text{tf}(w, s)$, where we boost the score of unigrams $w$ that appear in the first sentence of the article as $b(w)$. In the generic summarization $b(w) = 2$, for comparative summarization $b(w) = 3$ – as used by Gillick et al. (2009). Unigrams that do not appear in the first sentence of the article have $b(w) = 1$.

**lexrank** The LexRank score (Erkan and Radev, 2004) computed on the bigrams' cosine similarity.

For the TAC datasets, we additionally use:

**par_start**: An indicator whether the sentence begins a paragraph. This is provided by preprocessing pipeline from ICSI (Gillick et al., 2009).

**qsim**: The fraction of topic description unigrams present in each sentence; these topic descriptions are only available for TAC.

### 5.4 Oracle Extraction

Both DUC and TAC provide four human written summaries for each topic. Since our goal is extractive summarization with supervised training, we need to know which sentences in the articles could be used to construct the summaries in the training set. The article sentences that best match the abstractive summaries are called the *oracles* ($S^{\text{t}}$).

**Algorithm 1** Oracle extraction

**function** EXTRACTORACLE($\alpha, V^{\text{t}}, H^{\text{t}}, r, L$)
  $S^{\text{t}} \leftarrow \emptyset$
  **while** $\sum_{s \in S^{\text{t}}} \text{len}(s) \leq L$ **do**
    $s^* \leftarrow \underset{s \in V^{\text{t}} \setminus S^{\text{t}}}{\text{argmax}} \frac{\alpha(S^{\text{t}} \cup \{s\}, H^{\text{t}}) - \alpha(S^{\text{t}}, H^{\text{t}})}{\text{len}(s)^r}$
    $S^{\text{t}} \leftarrow S^{\text{t}} \cup \{s^*\}$
  **return** $S^t$

Our extraction algorithm (Algorithm 1), is inspired by Liu and Lapata (2019). We greedily select sentences ($s$) which provide the maximum gain in extraction score $\alpha(S^{\text{t}}, H^{\text{t}})$ against the human summaries ($H^{\text{t}}$) until a word budget ($L$) is reached. We only include sentences between 8 to 55 words as suggested by Yasunaga et al. (2017), and set a budget of 104 words to ensure our oracle summaries are within $100 \pm 4$ words, consistent with the evaluation ( §5.6).

In contrast to Liu and Lapata (2019) which uses only ROUGE-2 recall score (Lin, 2004), our method balances both ROUGE-1 and ROUGE-2 recall scores using the harmonic mean and explicitly accounts for sentence length. Grid search on the validation sets shows that the optimal value for $r$ is 0.4 across different datasets and summarization tasks. As reported in Table 1, on average our method produces oracles consisting of more sentences and with higher ROUGE-1 and ROUGE-2 scores compared to oracles from Liu and Lapata (2019). This is consistent across all datasets.

### 5.5 Implementation Details

Supervised variants use an $\ell_2$ regularized log-linear model of importance (§4.2) trained using the oracles (§5.4) as ground truth. We selected the number of training epochs using 5-fold cross validation. We then tune the other hyperparameters on the training set. The hyperparameters of the generic summarization task are: $\gamma$, a parameter of the kernel; $\beta$, the $\ell_2$ regularization weight for the log-linear importance function; and $r$, which defines the length dependent scaling factor in greedy selection (Lin and Bilmes, 2010). The comparative objective (4.6) has an additional hyperparameter $\lambda$, which controls the comparativeness. More implementation details are provided in the Appendix. We will make implementation publicly available[4].

[4] github.com/computationalmedia/supmmd

### 5.6 Evaluation Settings

To evaluate our methods we use the ROUGE (Lin, 2004) metric, the *de facto* choice for evaluating both generic summarization (Hong and Nenkova, 2014; Cho et al., 2019; Yasunaga et al., 2017; Kulesza and Taskar, 2012), and update summarization (Varma et al., 2009; Gillick and Favre, 2009; Zhang et al., 2009; Li et al., 2009). ROUGE metrics have been shown to correlate with human judgments (Lin, 2004) in generic summarization task. Our recent work (Bista et al., 2019) show that human judgments are consistent with the automatic metrics for evaluating comparative summaries.

Both DUC and TAC evaluations use the first 100 words of the generated summary. Our DUC-2004 evaluation setup mirrors (Hong et al., 2014). This allow us to compare performance with the state-of-the-art methods they reported and other works also evaluated using this setup[5]. As is standard for the DUC-2004 datasets, we report ROUGE-1 and ROUGE-2 recall scores.

For TAC-2009 datasets (both set A and B) , we adopt the evaluation settings from the TAC-2009 competition[6] so we can compare against the three best performing systems in the competition[7]. As is standard for the TAC-2009 dataset, we report ROUGE-2 and ROUGE-SU4 recall scores.

### 5.7 Baselines

**DUC-2004**: We select the top performing methods from a recent benchmark paper (Hong et al., 2014) to serve as baselines and report ROUGE scores from the benchmark paper. They are:
*ICSI*- an integer linear programming method that maximizes coverage (Gillick et al., 2009),
*DPP*- a determinantal point process method that learns sentence quality and maximizes diversity (Kulesza and Taskar, 2012),
*Submodular*- a method based on a learned mixture of submodular functions (Lin and Bilmes, 2012),
*OCCAMS_V*- a method base on topic modeling (Conroy et al., 2013),
*Regsum*- a method that focuses on learning word importance (Hong and Nenkova, 2014),
*Lexrank*- a popular graph based sentence scoring method (Erkan and Radev, 2004).

We also include recent deep learning methods evaluated in same setup as Hong et al. (2014) and report ROUGE scores from the individual papers:
*DPPSim* - an extension to the DPP model which learns the sentence-sentence similarity using a capsule network (Cho et al., 2019),
*HiMAP*- a recurrent neural model that employs a modified pointer-generator component (Fabbri et al., 2019), and
*GRU+GCN* - a model that uses a graph convolution network combined with a recurrent neural network to learn sentence saliency (Yasunaga et al., 2017).

**TAC-2009**: As baselines for the TAC-2009 dataset we use the top three systems in the TAC-2009 competition for each task, resulting four systems altogether. To the best of our knowledge these systems are the current state-of-the-art. We report the ROUGE scores from the competition. The systems are:
*ICSI*- with two variants: *Sys.34* uses integer linear programming to maximize coverage of concepts (Gillick et al., 2009), and *Sys.40*, which additionally uses sentence compression to generate new candidate sentences,
*IIT*- uses a support vector regressor to predict sentence ROUGE scores (Varma et al., 2009),
*ICTCAS*- a temporal content filtering method (Zhang et al., 2009), and
*ICL*- a manifold ranking based method (Li et al., 2009).

## 6 Experimental Results

We compare our methods with the baselines on the DUC-2004, TAC-2009-A and TAC-2009-B datasets. We present several variants of our method to analyze the effects of different components and modeling choices. We report the performance of unsupervised MMD (*UnsupMMD*) which does not explicitly consider sentence importance. For our supervised method *SupMMD*, we report the performance with a bigram kernel (*SupMMD*) and combined kernels (*SupMMD + MKL*). We also evaluated the impact of our oracle extraction method by replacing it with the extraction method suggested by Liu and Lapata (2019) in *SupMMD + alt oracles*. Meanwhile, *SupMMD + MKL + compress* presents the result of applying sentence compression (Gillick et al., 2009) to our model.

### 6.1 Generic Summarization

The performance of our methods on the DUC-2004 generic summarization task are shown in Table 2. On the DUC-2004 dataset all *SupMMD* variants

[5] ROUGE 1.5.5 with args -n 4 -m -a -l 100 -x -c 95 -r 1000 -f A -p 0.5 -t 0
[6] tac.nist.gov/2009/Summarization
[7] args -n 4 -w 1.2 -m -2 4 -u -c 95 -r 1000 -f A -p 0.5 -t 0 -a -l 100

| DUC-2004 | R1 | R2 |
|---|---|---|
| *ICSI* (Gillick et al., 2009) | 38.41 | 9.78 |
| *DPP* (Kulesza and Taskar, 2012) | **39.79** | 9.62 |
| *Submodular* (Lin and Bilmes, 2012) | 39.18 | 9.35 |
| *OCCAMS_V* (Conroy et al., 2013) | 38.50 | 9.76 |
| *Regsum* (Hong and Nenkova, 2014) | 38.57 | 9.75 |
| *Lexrank* Erkan and Radev (2004) | 35.95 | 7.47 |
| *DPP-Sim* (Cho et al., 2019) | 39.35 | 10.14 |
| *HiMAP* (Fabbri et al., 2019) | 35.78 | 8.90 |
| *GRU+GCN* (Yasunaga et al., 2017) | 38.23 | 9.48 |
| UnsupMMD | 35.73 | 7.76 |
| SupMMD (alt oracle) | 39.02 | 10.22 |
| SupMMD | 39.36 | 10.31 |
| SupMMD + MKL + compress | *39.63* | *10.50* |
| SupMMD + MKL | 39.27 | **10.54** |

Table 2: Results on DUC-2004 generic multi-document summarization task.

| TAC-2009-A | R2 | RSU4 |
|---|---|---|
| *ICSI*(Sys.34) (Gillick et al., 2009) | 12.10 | *15.09* |
| *ICSI*(Sys.40) (Gillick et al., 2009) | *12.16* | 15.03 |
| *IIIT*(Sys.35) (Varma et al., 2009) | 10.89 | 14.49 |
| *ICTCAS*(Sys.45) (Zhang et al., 2009) | 10.64 | 13.99 |
| UnsupMMD | 8.35 | 11.75 |
| SupMMD (alt oracle) | 11.13 | 14.22 |
| SupMMD | 11.76 | 14.67 |
| SupMMD + MKL + compress | 12.02 | 15.02 |
| SupMMD + MKL | **12.33** | **15.19** |

Table 3: Results on TAC-2009 generic multi-document summarization task (TAC-2009 set A).

exceed the state-of-the-art, when evaluated with ROUGE-2, and perform similarly to the best existing methods when evaluated with ROUGE-1. Our best system *SupMMD + MKL* outperforms the previous best system (*ICSI*) on ROUGE-2 score by $+\mathbf{3.9}\%$. While the *DPP* baseline achieves the highest ROUGE-1 score on DUC-2004, it has a relatively low ROUGE-2 score which suggests it is optimized for unigram performance at the cost of bigram performance. *SupMMD + MKL* strikes a better balance scoring the best in ROUGE-2 and second best in ROUGE-1. On the TAC-2009 generic summarization task in Table 3 our *SupMMD + MKL* model outperforms the state-of-the-art ICSI model on both ROUGE-2 and ROUGE-SU4. Specifically, *SupMMD + MKL* scores 12.33 in ROUGE-2 while the best ICSI variant scores 12.16 in ROUGE-2.

**Supervised Modeling**: Models using supervised training to identify important sentences substantially outperform the unsupervised method *UnsupMMD*. In fact, *UnsupMMD* is the lowest scoring method across all metrics and datasets. This strongly indicates that a degree of supervision is essential to perform well in this task, and that the importance function is a suitable way to adapt the *UnsupMMD* model to supervised training. Moreover, we observe a strong correlation between the the relative position of a sentence and the score given by *SupMMD*. This observation is consistent with previous works (Kedzie et al., 2018), and demonstrates that *SupMMD* has learned to use the surface features to capture salience. Further details of feature correlations are provided in the Appendix.

**Oracle extraction**: Our oracle extraction technique for transforming abstractive training data to extractive training data helps *SupMMD* methods achieve higher ROUGE performance. An alternative technique developed by Liu and Lapata (2019) and implemented in *SupMMD (alt oracle)* gives lower performance than our technique. For example, on DUC-2004 *SupMMD (alt oracle)* has a ROUGE-1 of 39.02 and ROUGE-2 of 10.22, while *SupMMD* has a ROUGE-1 of 39.36 and a ROUGE-2 of 10.31. Thus, the advantages of our proposed oracle extraction method are substantial and consistent across multiple datasets and evaluation metrics.

**Multiple Kernel Learning**: We observe that combining multiple kernels helps the performance of *SupMMD* models on the generic summarization task. *SupMMD + MKL* which combines both bigram and entity kernels has a ROUGE-2 of 10.54 on DUC-2004, while *SupMMD* only uses the bigrams kernel and scores 10.31 in ROUGE-2. Multiple kernels show even clearer gains in the TAC-2009-A dataset.

**Sentence compression** incorporated into the post-processing steps of *SupMMD + MKL + compress* does not clearly improve the results over *SupMMD + MKL*. On TAC-2009-A, compression clearly reduces performance, and on DUC-2004 *SupMMD + MKL + compress* has a higher ROUGE-1 score but a lower ROUGE-2 score than *SupMMD + MKL*. Incorporating compression into the summarization pipeline is an appealing direction for future work.

### 6.2 Comparative Summarization

The results for the comparative summarization task on the TAC-2009-B dataset are shown in Table 4. Our supervised MMD variants *SupMMD* and *SupMMD + MKL* both outperform the state-of-the-art baseline ICSI in ROUGE-SU4 but fall short in ROUGE-2. It would be hard to claim that either

method is superior in this instance; however, it does show that *SupMMD* – which uses a substantially different approach to that of *ICSI* – provides an alternative state-of-the-art. Thus *SupMMD* further maps out the set of techniques that are useful for comparative summarization. As per the generic summarization task, both our supervised training method and oracle extraction method are essential for achieving good performance in ROUGE-2 and ROUGE-SU4. We also identify sentence position and btfisf as important features for sentence salience (see the appendix).

**Multiple kernels** as in *SupMMD + MKL* has relatively little effect, reducing the ROUGE-2 score to 10.24 from the slightly higher 10.28 achieved by *SupMMD*. A similar small decrease is seen for ROUGE-SU4. Manual inspection shows that the summaries from *SupMMD* and *SupMMD + MKL* methods are largely identical with differences primarily on topic D0908, which covers political movements in Nepal. The key entities in this topic are not resolved accurately by DBPedia Spotlight, contributing additional noise and affecting the MKL approach.

**Model variants**: We have tested an additional variant of our model for comparative summarization, *SupMMD*$^2$, which defines two different importance functions: one for each of the two document sets - A and B (See §4.4 for details). In contrast, *SupMMD* has a single importance function shared between document sets, i.e., in Equation (4.6), $\theta_A = \theta_B$. *SupMMD*$^2$ performed substantially worse than *SupMMD* in both metrics, for example, *SupMMD* has a ROUGE-2 of 10.28 while *SupMMD*$^2$ has a ROUGE-2 of 9.94. We conjecture that a single importance function performs better when training data is relatively scarce because it reduces the number of parameters and simplifies the learning problem. Techniques for tying together the parameters for both importance functions, such as with a hierarchical Bayesian model, are left as future work.

## 7 Conclusions

In this work, we present SupMMD, a novel technique for update summarization based on the *maximum mean discrepancy*. SupMMD combines supervised learning for salience, and unsupervised learning for coverage and diversity. Further, we adapt multiple kernel learning to exploit multiple sources of similarity (e.g., text features and knowledge based concepts). We show the efficacy of SupMMD in both generic and update summarization tasks on two standard datasets, when compared to the existing approaches. We also show that the importance model we introduce on top of our existing unsupervised MMD (Bista et al., 2019) improves the summarization performance substantially on both generic and comparative summarization tasks.

For future work, we leave the task of incorporating embeddings features such as BERT (Devlin et al., 2019), and evaluating with large generic multi-document summarization dataset MultiNews (Fabbri et al., 2019).

| **TAC-2009-B** | **R2** | **RSU4** |
|---|---|---|
| *ICSI* (Sys.34) (Gillick et al., 2009) | **10.39** | 13.85 |
| *ICSI* (Sys.40) (Gillick et al., 2009) | *10.37* | 13.97 |
| *IIIT* (Sys.35) (Varma et al., 2009) | 10.10 | 13.84 |
| *ICL* (Sys.24) (Li et al., 2009) | 9.62 | 13.52 |
| UnsupMMD | 7.20 | 11.29 |
| SupMMD (alt oracle) | 10.06 | 13.86 |
| SupMMD$^2$ | 9.94 | 13.76 |
| SupMMD | *10.28* | **14.09** |
| SupMMD + MKL + compress | 10.25 | 13.91 |
| SupMMD + MKL | 10.24 | *14.05* |

Table 4: Results on TAC-2009 comparative multi-document summarization task (TAC-2009 set B).

## Acknowledgments

This work is supported in part by Data to Decisions CRC and ARC Discovery Project DP180101985. This research is also supported by use of the NeCTAR Research Cloud, a collaborative Australian research platform supported by the National Collaborative Research Infrastructure Strategy. We thank Minjeong Shin for helpful feedback and suggestions.

## References

Rasim M. Alguliev, Ramiz M. Aliguliyev, Makrufa S. Hajirahimova, and Chingiz A. Mehdiyev. 2011. Mcmr: Maximum coverage and minimum redundant text summarization model. *Expert Systems with Applications*, 38(12):14514 – 14522.

Steven Bird. 2006. NLTK: The Natural Language Toolkit. In *Proceedings of the COLING/ACL 2006 Interactive Presentation Sessions*, pages 69–72, Sydney, Australia. Association for Computational Linguistics.

Umanga Bista, Alexander Patrick Mathews, Minjeong Shin, Aditya Krishna Menon, and Lexing Xie. 2019.

Comparative document summarisation via classification. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*, pages 20–28. AAAI Press.

Ziqiang Cao, Furu Wei, Li Dong, Sujian Li, and Ming Zhou. 2015. Ranking with recursive neural networks and its application to multi-document summarization. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, AAAI'15, page 2153–2159. AAAI Press.

Sangwoo Cho, Logan Lebanoff, Hassan Foroosh, and Fei Liu. 2019. Improving the similarity measure of determinantal point processes for extractive multi-document summarization. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1027–1038, Florence, Italy. Association for Computational Linguistics.

John Conroy, Sashka T. Davis, Jeff Kubina, Yi-Kai Liu, Dianne P. O'Leary, and Judith D. Schlesinger. 2013. Multilingual summarization: Dimensionality reduction and a step towards optimal term coverage. In *Proceedings of the MultiLing 2013 Workshop on Multilingual Multi-document Summarization*, pages 55–63, Sofia, Bulgaria. Association for Computational Linguistics.

J.B. Conway. 1990. *A course in functional analysis*, second edition, volume 96 of *Graduate Texts in Mathematics*. Springer-Verlag, New York.

Corinna Cortes, Mehryar Mohri, and Afshin Rostamizadeh. 2010. Two-stage learning kernel algorithms. In *Proceedings of the 27th Annual International Conference on Machine Learning (ICML 2010)*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Gaël Dias, Elsa Alves, and José Gabriel Pereira Lopes. 2007. Topic segmentation algorithms for text summarization and passage retrieval: An exhaustive evaluation. In *Proceedings of the 22nd National Conference on Artificial Intelligence - Volume 2*, AAAI'07, page 1334–1339. AAAI Press.

Günes Erkan and Dragomir R. Radev. 2004. Lexrank: Graph-based lexical centrality as salience in text summarization. *J. Artif. Int. Res.*, 22(1):457–479.

Alexander Fabbri, Irene Li, Tianwei She, Suyi Li, and Dragomir Radev. 2019. Multi-news: A large-scale multi-document summarization dataset and abstractive hierarchical model. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1074–1084, Florence, Italy. Association for Computational Linguistics.

Dan Gillick and Benoit Favre. 2009. A scalable global model for summarization. In *Proceedings of the Workshop on Integer Linear Programming for Natural Language Processing*, pages 10–18, Boulder, Colorado. Association for Computational Linguistics.

Daniel Gillick, Benoit Favre, Dilek Hakkani-Tür, Bernd Bohnet, Yang Liu, and Shasha Xie. 2009. The icsi/utd summarization system at tac 2009. In *TAC*.

Jade Goldstein, Mark Kantrowitz, Vibhu Mittal, and Jaime Carbonell. 1999. Summarizing text documents: Sentence selection and evaluation metrics. In *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '99, page 121–128, New York, NY, USA. Association for Computing Machinery.

Arthur Gretton, Karsten M. Borgwardt, Malte J. Rasch, Bernhard Schölkopf, and Alexander Smola. 2012. A kernel two-sample test. 13(null):723–773.

Aria Haghighi and Lucy Vanderwende. 2009. Exploring content models for multi-document summarization. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 362–370, Boulder, Colorado. Association for Computational Linguistics.

Kai Hong, John Conroy, Benoit Favre, Alex Kulesza, Hui Lin, and Ani Nenkova. 2014. A repository of state of the art and competitive baseline summaries for generic news summarization. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC-2014)*, pages 1608–1616, Reykjavik, Iceland. European Languages Resources Association (ELRA).

Kai Hong and Ani Nenkova. 2014. Improving the estimation of word importance for news multi-document summarization. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 712–721, Gothenburg, Sweden. Association for Computational Linguistics.

Chris Kedzie, Kathleen McKeown, and Hal Daumé III. 2018. Content selection in deep learning models of summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1818–1828, Brussels, Belgium. Association for Computational Linguistics.

Been Kim, Rajiv Khanna, and Oluwasanmi Koyejo. 2016. Examples are not enough, learn to criticize! criticism for interpretability. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, NIPS'16, page 2288–2296, Red Hook, NY, USA. Curran Associates Inc.

Alex Kulesza and Ben Taskar. 2012. *Determinantal Point Processes for Machine Learning*. Now Publishers Inc., Hanover, MA, USA.

Sujian Li, Wei Wang, and Yongwei Zhang. 2009. Tac 2009 update summarization of icl. In *TAC*.

Yujia Li, Kevin Swersky, and Richard Zemel. 2015. Generative moment matching networks. In *Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37*, ICML'15, page 1718–1727. JMLR.org.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Hui Lin and Jeff Bilmes. 2010. Multi-document summarization via budgeted maximization of submodular functions. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 912–920, Los Angeles, California. Association for Computational Linguistics.

Hui Lin and Jeff Bilmes. 2011. A class of submodular functions for document summarization. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 510–520, Portland, Oregon, USA. Association for Computational Linguistics.

Hui Lin and Jeff Bilmes. 2012. Learning mixtures of submodular shells with application to document summarization. UAI'12, page 479–490, Arlington, Virginia, USA. AUAI Press.

Dong C Liu and Jorge Nocedal. 1989. On the limited memory bfgs method for large scale optimization. *Mathematical programming*, 45(1-3):503–528.

Yang Liu and Mirella Lapata. 2019. Text summarization with pretrained encoders. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3730–3740, Hong Kong, China. Association for Computational Linguistics.

Pablo N. Mendes, Max Jakob, Andrés García-Silva, and Christian Bizer. 2011. Dbpedia spotlight: Shedding light on the web of documents. In *Proceedings of the 7th International Conference on Semantic Systems*, I-Semantics '11, page 1–8, New York, NY, USA. Association for Computing Machinery.

Krikamol Muandet, Kenji Fukumizu, Bharath Sriperumbudur, Bernhard Schölkopf, et al. 2017. Kernel mean embedding of distributions: A review and beyond. *Foundations and Trends® in Machine Learning*, 10(1-2):1–141.

Ramesh Nallapati, Feifei Zhai, and Bowen Zhou. 2017. Summarunner: A recurrent neural network based sequence model for extractive summarization of documents. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, AAAI'17, page 3075–3081. AAAI Press.

George L Nemhauser, Laurence A Wolsey, and Marshall L Fisher. 1978. An analysis of approximations for maximizing submodular set functions—i. *Mathematical Programming*, 14(1):265–294.

Ani Nenkova and Lucy Vanderwende. 2005. The impact of frequency on summarization. *Microsoft Research, Redmond, Washington, Tech. Rep. MSR-TR-2005*, 101.

Joel Larocca Neto, Alexandre D Santos, Celso AA Kaestner, Neto Alexandre, D Santos, et al. 2000. Document clustering and text summarization.

Ganapati P Patil and Calyampudi R Rao. 1978. Weighted distributions and size-biased sampling with applications to wildlife populations and human families. *Biometrics*, pages 179–189.

Dragomir R Radev, Hongyan Jing, Małgorzata Styś, and Daniel Tam. 2004. Centroid-based summarization of multiple documents. *Information Processing & Management*, 40(6):919–938.

Bharath K. Sriperumbudur, Kenji Fukumizu, Arthur Gretton, Gert R. G. Lanckriet, and Bernhard Schölkopf. 2009. Kernel choice and classifiability for rkhs embeddings of probability distributions. In *Proceedings of the 22nd International Conference on Neural Information Processing Systems*, NIPS'09, page 1750–1758, Red Hook, NY, USA. Curran Associates Inc.

Bharath K Sriperumbudur, Arthur Gretton, Kenji Fukumizu, Bernhard Schölkopf, and Gert RG Lanckriet. 2010. Hilbert space embeddings and metrics on probability measures. *Journal of Machine Learning Research*, 11(Apr):1517–1561.

Ingo Steinwart. 2001. On the influence of the kernel on the consistency of support vector machines. *Journal of machine learning research*, 2(Nov):67–93.

Vasudeva Varma, Prasad Pingali, Rahul Katragadda, Sai Krishna, Surya Ganesh, Kiran Sarvabhotla, Harish Garapati, Hareen Gopisetty, Vijay Bharath Reddy, Kranthi Reddy, et al. 2009. Iiit hyderabad at tac 2009. In *TAC*.

Xiaojun Wan, Jianwu Yang, and Jianguo Xiao. 2007. Towards an iterative reinforcement approach for simultaneous document summarization and keyword

extraction. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 552–559, Prague, Czech Republic. Association for Computational Linguistics.

Michihiro Yasunaga, Rui Zhang, Kshitijh Meelu, Ayush Pareek, Krishnan Srinivasan, and Dragomir Radev. 2017. Graph-based neural multi-document summarization. In *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)*, pages 452–462, Vancouver, Canada. Association for Computational Linguistics.

Manzil Zaheer, Sashank Reddi, Devendra Sachan, Satyen Kale, and Sanjiv Kumar. 2018. Adaptive methods for nonconvex optimization. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems 31*, pages 9793–9803. Curran Associates, Inc.

Jin Zhang, Pan Du, Hongbo Xu, and Xueqi Cheng. 2009. Ictgrasper at TAC2009: temporal preferred update summarization. In *Proceedings of the Second Text Analysis Conference, TAC 2009, Gaithersburg, Maryland, USA, November 16-17, 2009*.

# A Background Theory on Kernels and MMD

In this section, we provide a brief overview of kernels and Maximum mean Discrepancy (MMD). For a detailed overview, we refer readers to to Muandet et al. (2017) and Gretton et al. (2012) from which this brief overview is taken.

## A.1 Positive Definite Kernels and Kernel Trick

**Definition A.1.** A function $k : \mathcal{X} \times \mathcal{X} \mapsto \mathbb{R}$ is called positive definite kernel if it is symmetric, i.e. $\forall_{x,y \in \mathcal{X}}\ k(x,y) = k(y,x)$ and gram matrix is positive definite, i.e. $\forall_{n \in \mathbb{N}}\ \forall_{c_1,c_2,..c_n \in \mathbb{R}}\ \sum_{i,j=1}^{n} c_i c_j k(x_i, x_j) \geq 0$.

**Theorem A.1.** If a kernel is positive definite, there exists a feature map $\phi : \mathcal{X} \mapsto \mathcal{H}$ such that $\forall_{x,y \in \mathcal{X}}\ k(x,y) = \langle \phi(x),\ \phi(y) \rangle_{\mathcal{H}}$.

This is known as the kernel trick in machine learning. The feature space $\mathcal{H}$ is called a Reproducing Kernel Hilbert Space (RKHS), and the kernel $k$ is also known as reproducing kernel.

## A.2 Reproducing Kernel Hilbert Space

**Definition A.2.** An RKHS is a Hilbert space of functions where all function evaluations are bounded, i.e. $\forall_{x \in \mathcal{X}}\ \forall_{h \in \mathcal{H}}\ \exists_{c>0}\ |h(x)| \leq c\|h\|_{\mathcal{H}}$.

In an RKHS, function evaluation $h(x) = \langle h,\ \phi(x) \rangle_{\mathcal{H}}$, where $\phi : \mathcal{X} \mapsto \mathcal{H}$ are canonical feature map associated with RKHS $\mathcal{H}$, and $\phi(x) = k(.,x)$. A RKHS is fully characterized by its reproducing kernel $k$, or a positive definite kernel $k$ uniquely determines a RKHS and vice versa . Hence, $\mathbb{E}_p[h(x)] = \langle h,\ \mathbb{E}_p[\phi(x)] \rangle_{\mathcal{H}}$, which is known as the Riesz representer theorem (Conway, 1990).

## A.3 More on MMD

Recall that $\mathcal{F}$ is a class of RKHS functions within the unit ball, i.e. $h \in \mathcal{H}, \|h\|_{\mathcal{H}} \leq 1$. Suppose $\mathcal{H}$ admits a *feature map* $\phi \colon \mathcal{X} \mapsto \mathcal{H}$. Then, per Gretton et al. (2012), we may solve the supremum in Equation 3.3 as

$$\mathrm{MMD}_{\mathcal{F}}(p,q) = \|\mathbb{E}_p \phi(x) - \mathbb{E}_q \phi(y)\|_{\mathcal{H}}\,. \quad \text{(A.1)}$$

Hence, MMD is computed as the distance between the *mean feature embeddings* under each distribution, for a suitable kernel-based feature space (Gretton et al., 2012).

Eq. (A.1) involves explicitly evaluating the arbitrarily high-dimensional features. Instead, the *kernel trick* allows efficient computation of $\mathrm{MMD}^2_{\mathcal{F}}(p,q)$ by evaluating just pairwise kernels. Supposing $\mathcal{H}$ has induced kernel $k$, we have

$$\mathrm{MMD}^2_{\mathcal{F}}(p,q) = \mathop{\mathbb{E}}_{x,x' \sim p}\left[k(x,x')\right] + \mathop{\mathbb{E}}_{y,y' \sim q}\left[k(y,y')\right] - 2\mathop{\mathbb{E}}_{x \sim p, y \sim q}\left[k(x,y)\right]. \quad \text{(A.2)}$$

## A.4 Characteristic Kernel

For a distribution $p$, and kernel with feature map $\phi \colon \mathcal{X} \mapsto \mathcal{H}$, the *kernel mean map* is

$$\mu_p = \mathbb{E}_{x \sim p}\left[\phi(x)\right].$$

A kernel $k$ is characteristic if the map $\mu : p \mapsto \mu_p$ is injective. A characteristic kernel ensures MMD is 0 if and only if $p = q$, i.e., no information is lost in mapping the distribution into the RKHS (Muandet et al., 2017).

Examples of characteristic kernels for $\mathbb{R}^d$ include the Gaussian kernel ( $k(\mathbf{x},\mathbf{y}) = \exp\left(-\gamma\|\mathbf{x}-\mathbf{y}\|_2^2\right),\ \gamma > 0$ ), and Laplace kernel ( $k(\mathbf{x},\mathbf{y}) = \exp(-\gamma\|\mathbf{x}-\mathbf{y}\|_1),\ \gamma > 0$ ) . MMD with the Gaussian kernel is equivalent to comparing all moments between two distributions (Li et al., 2015).

## B Proof of Lemma 4.1

The weighted MMD $\mathrm{MMD}_{\mathcal{F}}(p, q, \theta)$ (§4.1), where $\mathcal{F}$ contains functions $h : \mathcal{X} \mapsto \mathbb{R}$ within unit ball RKHS $\mathcal{H}$ ($\|h\|_{\mathcal{H}} \leq 1$) is defined as:

$$\sup_{h \in \mathcal{F}} \left( \mathop{\mathbb{E}}_{v \sim p} \left[ f^p_\theta(v) \cdot h(v) \right] - \mathop{\mathbb{E}}_{s \sim q} \left[ f^q_\theta(s) \cdot h(s) \right] \right)$$

Recall $f_\theta$ is a non-negative importance weighting function. Then, according to Patil and Rao (1978), the weighted probability density $\bar{p}_\theta$ of $p$ is:

$$\bar{p}_\theta(v) = \frac{f^p_\theta(v) \cdot p(v)}{\mathbb{E}_p[f^p_\theta(v)]}$$

and similarly $\bar{q}_\theta$ for $q$. Since we restrict $\mathbb{E}_p[f^p_\theta(v)] = 1$, and $\mathbb{E}_q[f^q_\theta(s)] = 1$, we have $\bar{p}_\theta(v) = f^p_\theta(v) \cdot p(v)$ and $\bar{q}_\theta(s) = f^q_\theta(s) \cdot q(s)$. Thus, the weighted MMD is

$$\begin{aligned} &\sup_{h \in \mathcal{F}} \left( \mathop{\mathbb{E}}_{v \sim \bar{p}_\theta} [h(v)] - \mathop{\mathbb{E}}_{s \sim \bar{q}_\theta} [h(s)] \right) \\ &= \sup_{||h||_{\mathcal{H}} \leq 1} \left( \mathop{\mathbb{E}}_{v \sim \bar{p}_\theta} [h(v)] - \mathop{\mathbb{E}}_{s \sim \bar{q}_\theta} [h(s)] \right) \end{aligned}$$

Since in an RKHS, $\mathbb{E}_p[h(x)] = \langle h,\ \mathbb{E}_p[\phi(x)] \rangle_{\mathcal{H}}$, this simplifies to:

$$\begin{aligned} &\sup_{||h||_{\mathcal{H}} \leq 1} \left\langle h,\ \mathop{\mathbb{E}}_{v \sim \bar{p}_\theta} [\phi(v)] - \mathop{\mathbb{E}}_{s \sim \bar{q}_\theta} [\phi(s)] \right\rangle_{\mathcal{H}} \\ &= \left\| \mathop{\mathbb{E}}_{v \sim \bar{p}_\theta} [\phi(v)] - \mathop{\mathbb{E}}_{s \sim \bar{q}_\theta} [\phi(s)] \right\|_{\mathcal{H}} \\ &= \left\| \mathop{\mathbb{E}}_{v \sim p} \left[ f^p_\theta(v) \cdot \phi(v) \right] - \mathop{\mathbb{E}}_{s \sim q} \left[ f^q_\theta(s) \cdot \phi(s) \right] \right\|_{\mathcal{H}}, \end{aligned}$$

where the penultimate step follows from the dual norm theorem[8]. The proof is similar to MMD in (Gretton et al., 2012).

## C Empirical estimate of $\mathrm{MMD}^2_{\mathcal{F}}(p, q, \theta)$

First, $\mathrm{MMD}^2_{\mathcal{F}}(p, q, \theta)$ can be expanded as:

$$\begin{aligned} &\left\| \mathop{\mathbb{E}}_{v \sim p} \left[ f^p_\theta(v) \cdot \phi(v) \right] - \mathop{\mathbb{E}}_{s \sim q} \left[ f^q_\theta(s) \cdot \phi(s) \right] \right\|^2_{\mathcal{H}} \\ &= \mathop{\mathbb{E}}_{v, v' \sim p} \left[ f^p_\theta(v) \cdot f^p_\theta(v') \cdot \langle \phi(v), \phi(v') \rangle_{\mathcal{H}} \right] \\ &\quad - 2 \cdot \mathop{\mathbb{E}}_{v \sim p, s \sim q} \left[ f^p_\theta(v) \cdot f^q_\theta(s) \cdot \langle \phi(v), \phi(s) \rangle_{\mathcal{H}} \right] \\ &\quad + \mathop{\mathbb{E}}_{s, s' \sim q} \left[ f^q_\theta(s) \cdot f^q_\theta(s') \cdot \langle \phi(s), \phi(s') \rangle_{\mathcal{H}} \right] \end{aligned}$$

[8] https://en.wikipedia.org/wiki/Dual_norm

Applying the kernel trick (A.2),

$$\begin{aligned} &= \mathop{\mathbb{E}}_{v, v' \sim p} \left[ f^p_\theta(v) \cdot f^p_\theta(v') \cdot k(v, v') \right] \\ &\quad - 2 \cdot \mathop{\mathbb{E}}_{v \sim p, s \sim q} \left[ f^p_\theta(v) \cdot f^q_\theta(s) \cdot k(v, s) \right] \\ &\quad + \mathop{\mathbb{E}}_{s, s' \sim q} \left[ f^q_\theta(s) \cdot f^q_\theta(s') \cdot k(s, s') \right] \end{aligned}$$

Our loss of generic summarization $\mathcal{L}^{\mathrm{t}}(V^{\mathrm{t}}, S^{\mathrm{t}}, \theta)$ is $\mathrm{MMD}^2_{\mathcal{F}}(V^{\mathrm{t}}, S^{\mathrm{t}}, \theta)$. Recalling $n_{\mathrm{t}} = |V^{\mathrm{t}}|$ and $m_{\mathrm{t}} = |S^{\mathrm{t}}|$:

$$\begin{aligned} \mathcal{L}^{\mathrm{t}} = &\frac{1}{n^2_{\mathrm{t}}} \sum_{v, v'} f^{n_{\mathrm{t}}}_\theta(v) \cdot f^{n_{\mathrm{t}}}_\theta(v') \cdot k(v, v') \\ &- \frac{2}{n_{\mathrm{t}} \cdot m_{\mathrm{t}}} \sum_{v, s} f^{n_{\mathrm{t}}}_\theta(v) \cdot f^{m_{\mathrm{t}}}_\theta(s) \cdot k(v, s) \\ &+ \frac{1}{m^2_{\mathrm{t}}} \sum_{s, s'} f^{m_{\mathrm{t}}}_\theta(s) \cdot f^{m_{\mathrm{t}}}_\theta(s') \cdot k(s, s') \end{aligned}$$

## D Training details

We train generic summarization model with full batch LBFGS (Liu and Nocedal, 1989) with learning rate 0.005. We train comparative summarization model using Yogi optimizer (Zaheer et al., 2018), with mini batch size of 8 topics, learning rate 0.002, and decreasing the learning rate by half every 20 epochs. We choose the number of training epochs by validating across 5 folds with early stopping. We use patience of 20 epochs for early stopping with LBFGS optimizer and 50 epochs with Yogi optimizer. We tune the other hyperparameters on the training set, and the optimal hyperparameters of best model (SupMMD + MKL) and searched space are shown in Table 5. The kernel combination weights $\mathbf{w}$ (§4.5) are also shown in Table 5. The kernel combination weights ($\mathbf{w}$) are written in order: unigrams, bigrams and entities.

| hyp. | DUC-2003 | TAC-2008-A | TAC-2009-B |
|---|---|---|---|
| $\gamma$ | 2.5[1-4] | 4.5[2-6] | 2.2[1-3] |
| $\beta$ | 0.04[.02-.16] | 0.08[.02-.16] | 0.02[.01-.16] |
| $\lambda$ | - | - | 0.5[.25-.625] |
| $r$ | 0.001[0-.01] | 0.01[-0.01] | 0.01[-0.01] |
| epoch | 64 | 53 | 94 |
| $\mathbf{w}$ | [.0, .968, .032] | [.01, .97, .02] | [.014, .98, .006] |

Table 5: Optimal hyperparameters, their search space and MKL combination weights on each dataset.

## E Additional results

In this section we provide some additional results.

| feature | DUC2004 SupMMD | DUC2004 LexRank | TAC2009-A SupMMD | TAC2009-A LexRank | TAC2009-B SupMMD | TAC2009-B LexRank |
|---|---|---|---|---|---|---|
| position | 0.34 | 0.16 | 0.32 | 0.18 | 0.44 | 0.22 |
| tfisf | 0.07 | 0.38 | 0.22 | 0.37 | 0.01 | 0.36 |
| btfisf | 0.30 | 0.52 | 0.48 | 0.53 | 0.46 | 0.57 |
| #words | 0.0 | 0.35 | 0.08 | 0.33 | -0.15 | 0.31 |
| #nouns | 0.15 | 0.43 | 0.27 | 0.41 | 0.08 | 0.40 |

Table 6: Correlation of some features with sentence scores from SupMMD and Lexrank eigenvector centrality.

| method | set A | set B |
|---|---|---|
| ICSI | **A fourth day** of thrashing thunderstorms began to take a heavier toll on southern California with at least **three deaths** blamed on the rain, as flooding and mudslides forced road closures and emergency crews carried out harrowing rescue operations. Downtown Los Angeles has had more than 15 inches of rain since Jan. 1, more than its average rainfall for an entire year, including 2.6 inches, a record. Meteorologists say Southern California has not been hit by **this much rain in nearly 40 years**. The disaster was the latest caused by **rain and snow** that has battered California since Dec. 25. | Californians braced for **even more rain** as they struggled to recover from storms that have left at least **nine people dead**, triggered mudslides and tornadoes, and washed away roads and runways. The **record**, 38.18 inches (96.98 centimeters), was set in **1883-1884**. **Mudslides** forced Amtrak to suspend train service between Los Angeles and Santa Barbara through at least Thursday. A **winter storm** pummeled Southern California for the third straight day, claiming the lives of three people and raising fears of mudslides, even as homes around the region were evacuated. Staff Writers Rick Orlov and Lisa Mascaro contributed to this story. |
| SupMMD | Downtown Los Angeles has had more than **15 inches of rain** since Jan. 1, more than its average rainfall for an entire year, including 2.6 inches, a record. A **fourth day** of thrashing thunderstorms began to take a heavier toll on southern California with at least three deaths blamed on the rain, as **flooding and mudslides** forced road closures and emergency crews carried out harrowing rescue operations. The roads in Los Angeles County were equally frustrating. Part of a rain-saturated hillside gave way, sending a Mississippi-like **torrent of earth and trees** onto four blocks of this oceanfront town and killing two men. | Storms have caused **$52.5 million** (euro39.8 million) in damage to Los Angeles County roads and facilities since the beginning of the year. Multi-million-dollar homes collapsed and mudslides trapped residents in their homes as a heavy rains that have claimed three lives pelted Los Angeles for the fifth straight day. In scenes reminiscent of the aftermath of the Northridge Earthquake 11 years ago this month, Los Angeles area residents faced gridlocked freeways and roads Wednesday while cleanup crews cleared mud, rubble and debris left from a **two-week siege of rain**. A **record-shattering storm** slammed Southern California for a **sixth straight day** Tuesday, triggering **mudslides and tornadoes** and forcing more road closures, but forecasters predicted it would wane Wednesday before a new storm moves in Sunday night. |

Table 7: Example summaries of topic D0906, containing articles about "Rains and mudslides in Southern California".

### E.1 Correlation with rouge score

| dataset | ROUGE-2 SupMMD | ROUGE-2 LexRank | ROUGE-1 SupMMD | ROUGE-1 LexRank |
|---|---|---|---|---|
| TAC2009A | 0.590 | 0.555 | 0.571 | 0.543 |
| DUC2004 | 0.595 | 0.577 | 0.567 | 0.545 |

Table 8: Correlation of sentence importance scores with normalized sentence ROUGE scores.

We analyze the correlation between normalized ROUGE recall scores of the sentences and sentence scores from *SupMMD* and *Lexrank*. The normalized rouge score of each sentence is defined as $\overline{\text{ROUGE}}(s) = \frac{\text{ROUGE}(s)}{\#\text{words}(s)}$. As shown in Table 8, we find that *SupMMD* has slightly high correlation with sentence rouge scores. This suggests that *SupMMD* is better in capturing sentence importance for summarization.

### E.2 Feature correlations

We analyze the correlation between various surface features and sentence importance scores from *SupMMD* and *Lexrank* (Erkan and Radev, 2004). As shown in table 6, *SupMMD* has higher correlation with relative position, signifying the importance of position of sentence in summary sentences. *Lexrank* has higher correlations with the number of words, number of nouns and TFISF scores of the sentences, which is expected as *Lexrank* is an eigenvector centrality of sentence-sentence similarity matrix. This suggest *SupMMD* is able to learn that first few sentences are important in news summarization. Similar result is reported by Kedzie

et al. (2018), where they show first few sentences are important in creating summary of news articles.

### E.3 Example summary

We present the update summaries (Set A and B) of topic D0906, which contains articles about *"Rains and mudslides in Southern California"* in Table 7. We highlight few phrases in bold which could help us to identify the difference between set A and B. Summaries from *ICSI* and *SupMMD* methods suggest that set A contains articles describing events from earlier days of the disaster and set B contains articles from later stage of the disaster.