# Global and Local Models for Multi-Document Summarization

**Pradipto Das**
SUNY Buffalo
CSE Dept.
Buffalo, NY 14260
pdas3@buffalo.edu

**Rohini Srihari**
SUNY Buffalo
CSE Dept.
Buffalo, NY 14260
rohini@cedar.buffalo.edu

## Abstract

In this paper we study the effectiveness of combining corpus-level (global) tag-topic models and target document set level local models for multi-document summarization. Recently tag-topic models that exploit both word level annotation (e.g. named entity type) and/or document level metadata (e.g. words related to topic categories) have been proposed to model documents tagged from two different perspectives. We augment such models with an informative prior over document level latent topic proportions and conduct extensive experiments on multi-document summarization of newswire articles within the TAC 2010 and 2011 datasets. We train tag-topic models on the entire set of *documents* without knowledge of any document relevancy and fit each *sentence* of the documents in the relevant document set to the models. Combining the model likelihoods of these sentences and their similarities to bag-of-words derived from key terms in the relevant document set resulted in summaries that significantly outperformed robust baselines. By augmenting this model with local Rhetorical Structure trees of these sentences, we are able to select the salient spans of the sentences which are used to generate bulleted list summaries. We empirically show that the standard ROUGE SU4 scores of such summaries are comparable to those obtained from human generated counterparts.

## 1 Introduction

Guided multidocument summarization is the task of generating a summary of a collection of documents as an answer to an information need of a user. In the TAC 2010 and 2011 guided summarization tasks, these basic information needs are commonly expressed as very short query title strings together with an unified information model of categories and aspects which all summaries are expected to cover. In general, solutions for automatic text summarization is approached as a combination of several factors: the importance of sentences (which can be estimated from how often they are paraphrased across the collection), the redundancy between sentences (so as not to generate redundant summaries), and the readability of the produced summary. Because of its simplicity, most summarization systems currently used are extractive, i.e. they compose the output summary by combining sentences extracted from the original documents, which are sometimes modified through sentence rewriting or compression. However, experiments on human extractive summarization (Genest, 2010) show that even the best content-selection mechanism (e.g., a human summarizer) that is limited to pasting together sentences cannot achieve the same quality as fully manual summaries. The analysis of the rhetorical structure of texts has shown promise in the past for text summarization (Marcu, 1999) so we believe this direction should be further explored.

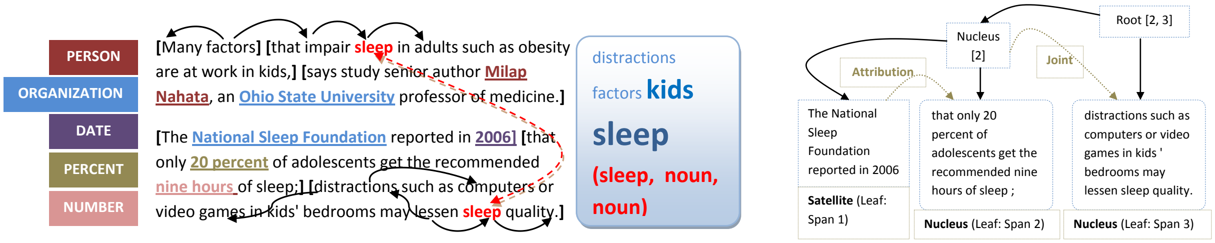
Current state of the art extractive query-focused summarization systems like CLASSY (Conroy, 2010) use similar techniques. Regarding sentence scoring, a very important aspect of all query-focused summarization systems is to model the importance of words in the sentences conditioned on the user's information need. Many systems, including CLASSY, derive a lexicon that best represents the the categorical concepts through the use of external sources like the internet. However, it was recently noted in (Conroy, 2010) that such lexicons may lower summarization performance due to topic drift. We show in this paper how simple models that are local to the docsets can be used to derive such lexicon automatically from the data at hand. We find that such automatically derived lexicon is very appropriate for categorizing documents in the TAC categories, and for summarizing the documents according to the guided summarization task definition.

Alternatively, unsupervised topic models like LDA(Blei, 2003) are very powerful data exploration techniques which can summarize data in the form of bag-of-word summaries where each bag holds semantically related items. Recent extensions of LDA-based models that use more structure in the representation of documents have also been proposed for generating more coherent and less redundant summaries, such as those in (Asli, 2011; Aria, 2009; Daume, 2006). These models use the collection and target document-specific distributions in order to distinguish between the general and specific topics in documents. In the context of summarization, this distinction is very similar to identifying signature terms(Lin, 2000) at multiple granularities in a corpus driven manner and weight sentences accordingly for inclusion into summaries. Since many of these signature terms happen to be Named Entities, it is often useful to use supervised methods to identify them and influence the topic modeling process instead. In this paper, one of the main reasons to choose multi-modal tag topic models(Das, 2011) was their ability to handle the word level annotations. The aspects of the categories concerning {Who, when, date, location} naturally ask for highlighting the text with Named Entities and that the discovery of latent topics should also be influenced by the presence of these entities.

In addition to the experiments already performed during the main TAC2011 Guided Summarization task competition in July 2011, we changed several aspects of our system to improve the summarization scores for the initial summarization task. Our recent experiments have the following implications: 1) We extend the class of topic models in (Das, 2011) that we used initially with informative prior over the document level topic proportions to drive better topic generation, better topic-event classification of documents and subsequent summarization. 2) We show that fitting the extended tag-topic models to sentences in target document sets produce better results for guided summarization. It is observed that without the knowledge of any relevant docset structure for use by the tag-topic models during training, the summaries formed out of such sentences easily parallel those from a very robust algebraic baseline summarization system which uses features based on the relevancy knowledge of documents. We present an approach to apply these topic models in multi-document summarization by combining sentence likelihoods from the extended tag-topic models with those from very granular yet simple target document set specific local models to vastly improve summarization performance. 3) Finally, we use sub-sentential spans automatically obtained from rhetorical parsers that implement Rhetorical Structure Theory(Mann, 1988) to create bulleted summaries for better readability and including more information in less space.

### 1.1 Motivation for Using Global Tag-Topic Models and Local Sentential Models

Our hypothesis is that there are a number of finer to coarser latent features in documents that can be very useful for the task of summarization. These include automatically discovered latent topic clusters, dependencies within sentential words, coherence structure, rhetorical structure, etc. and we wanted a model that captures several of these things and optimizes them jointly. Figure 1a shows two sentences from a sample text concerning sleep deprivation. The light blue rectangular bubble on the right contains words stylized in varying font sizes depending on their frequencies in the text. As in (Das, 2011), this bag-of-words can be looked upon as a document level perspective that provides a gist of the document in terms of salient words appearing most frequently. The frequency of words usually have considerable impact on final summaries(Nenkova, 2006). The text itself can be structured in many ways. In this example, each word either belongs to a particular Named Entity class or not. Here we show 5 such classes with the corresponding phrases in the text appropriately color coded and underlined. The


[Many factors] [that impair sleep in adults such as obesity are at work in kids,] [says study senior author Milap Nahata, an Ohio State University professor of medicine.]

[The National Sleep Foundation reported in 2006] [that only 20 percent of adolescents get the recommended nine hours of sleep;] [distractions such as computers or video games in kids' bedrooms may lessen sleep quality.]

PERSON
ORGANIZATION
DATE
PERCENT
NUMBER
distractions
factors kids
sleep
(sleep, noun, noun)
Root [2, 3]
Nucleus [2]
Attribution
Joint
The National Sleep Foundation reported in 2006
Satellite (Leaf: Span 1)
that only 20 percent of adolescents get the recommended nine hours of sleep ;
Nucleus (Leaf: Span 2)
distractions such as computers or video games in kids ' bedrooms may lessen sleep quality.
Nucleus (Leaf: Span 3)

(a) Word level and document level perspectives

(b) Rhetorical Structure Tree of the second sentence

Figure 1: An article on sleep deprivation showing two perspectives with some shallow and deep linguistic structures

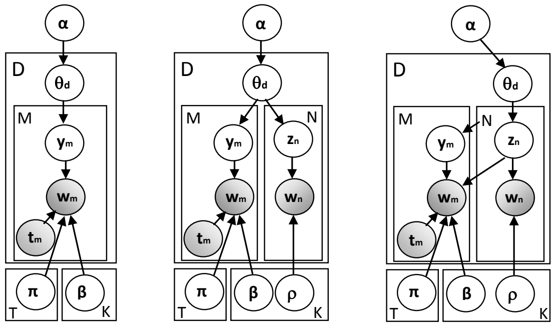
words not colored do not belong to any Named Entity class. The word "sleep" (connected by a dashed arrow) in **bold** font and colored red appears as a *noun* in the first and second sentences. To a reader the word "sleep" represented as the triplet **(sleep, noun, noun)** act as an important center of attention that signifies an event rather than an entity. This triple thus also helps strengthening the document level perspective. In the triplet the first element is the word that appears in both sentences, the second one is the role of the word in the first sentence and the third one — its role in the consecutive sentence. The black arcs show extremely fine-grained syntactic and semantic dependencies that exist between selected words. In our case study, an important observation was that salient high frequency verbs (i.e. verbs that do not fall into the category of standard English stopwords) across documents relevant to an information need identify the main events to a considerable degree. Here we become aware that something is being discussed around the concept of "impairment of sleep". If verbs like "impair" or "deprive" occur frequently across the docset, then we actually recover the query!

We observe that the central ideas in a document are often conveyed in written English through syllogisms. These logical inference constructs often lead to the propagation of certain important concepts similar in spirit to "centers of utterances" in (Grosz, 1995). The propagation of these centers, be they entities or otherwise, are a major contributor to the high frequency of certain open-class words in documents. Intuitively summarization is best described as an abstraction activity. Too much focus on intricate details of inter-word dependencies in sentences can lead to loss of context. In figure 1a, the WL annotations by Named Entity classes lead to the creation of the word level perspective as in (Das, 2011). Since the category labels of the manual topics indicated some natural events, we were able to see how much generalization power did the tag-topic models possess vis-a-vis simple feature sets to classify documents as belonging to the right event classes.

RST literature (Mann, 1988; Marcu, 2000) lays special emphasis on cue words or phrases which are sentence level connectives like "because", "nevertheless", "that", "but", "in spite of", parenthesis etc. and certain punctuations that serve primarily to indicate document structure or flow. In figure 1a, the square-bracketed textual extents represent such sub-sentential spans as recognized by cue words. RST emphasizes the fact that certain shallow processing of text in terms of analysis of cue phrases in combination with well-constrained mathematical models can be used to create valid rhetorical structure trees (RS-trees) of unconstrained natural language text. Rhetorical parsing allows a piece of text to be partitioned into non-overlapping segments called spans and then using rigorous formulations (first-order logic or otherwise) and training statistics to build a binary tree of the text where the leaves from left to right indicate elementary discourse units that are related in strict rhetorical sense. Any internal node signifies a relationship between its children i.e. the text extents only covered by the children. The spans of the text, as broken by identification of the cue phrases, are of two types - text spans that consume subsidiary information are called **satellites** and all others are called **nuclei**. All satellites are related to its corresponding nucleus through some valid rhetorical relation. Figure 1b shows the RS-tree of the second sentence in fig. 1a. The parent for spans 1 and 2 becomes a nucleus signifying that span 2 is more important. The root indicates that both spans 2 and 3 are jointly important and that they are related through the rhetorical relation of "Joint". In our paper, the RS trees of the sentences had been generated using the techniques used in (Soricut, 2003). We have slightly modified the accompanying software for (Soricut, 2003)[1], to incorporate minor modifications and bug fixes. The rhetorical relations that hold between different spans of text are the same as those used in (Soricut, 2003). We consider only the following relations to be useful for our purposes: {Background, Cause, Cause-Result, Comparison, Consequence, Contrast, Explanation and Temporal} to locally emphasize the aspects of the topic-categories that are more subtle and cannot be handled by Named Entity annotation. The words in the satellites corresponding the these rhetorical relationship classes are not used in any further WL annotation to be used in the tag-topic models but rather used in the local models as a criteria for inclusion into summary sentences. In fig. 1b, we can think of spans 2 and 3 as good summary spans from the second sentence because of a global or background topic focus, presence of topically salient numeric entities, relevance to the query and the importance of the spans. We are thus motivated to use both background topic models that look at the corpus as a whole and local docset and sentential models for the guided summarization problem.

## 2 The TagLDA and Tag$^2$LDA Models

In this section we augment the TagLDA model(Zhu, 2006) and the multimodal tag-topic models in (Das, 2011) with an asymmetric topic proportion prior. We use the procedure described in (Blei, 2003) and the references therein. The use of this prior was also motivated by the work in (Wallach, 2009). In other words we optimize the $K$ dimensional $\boldsymbol{\alpha}$ in figure 2 such that each dimension is treated differently. We rename the multi-modal METag$^2$LDA model and the correspondence corrMETag$^2$LDA models in (Das, 2011) as MPTag$^2$LDA model and corrMPTag$^2$LDA models respectively. We do so since the word generation probabilities are simply obtained by the product of two distributions - $\boldsymbol{\beta}$: the marginal topic-word distributions and $\boldsymbol{\pi}$: the marginal WL_tag-word distributions. The "M" and "P" stands for Multinomial (as in (Das, 2011)) and Product respectively. The MP Tag$^2$LDA model is shown in Fig. 2b and the corrMPTag$^2$LDA is shown in Fig. 2c. For a full generative story of these models, we refer the reader to (Das, 2011).



(a) TagLDA model

(b) MPTag$^2$LDA model

(c) corrMPTag$^2$LDA model

Figure 2: Graphical model representations of the tag-topic models used in modeling the corpus

To create the DL perspective from plain text documents to be used in the tag-topic models we used the top 5 most frequent non-stopwords in each document. All tokens were lemmatized prior to any usage. We also added top tf-idf terms per document until there were 20 such tokens at the document level. Following the example of the word "sleep" in fig. 1a, the same lemmatized words in consecutive sentences were extended with the the part-of-speech tag or the syntactic dependency labels (e.g. a triplet like (arrest, subject, noun)). These also became part of the DL perspective. This is similar in spirit as (Barzilay, 2005) but not restricted to Named Entities only. In this paper we refer to such a triplet as *coherence marker*. It was quite surprising to find that the intersection size between the 5 most frequent words and the set of all the first elements of such coherence markers is 3.5 on average per document even without co-reference resolution. This is a pretty good testament to the fact that the co-occurrence pattern in English text actually follows from coherence properties within textual units. We used Named Entity annotation classes as word level tags and a "Normal_Word" tag for all other words. All entities were automatically recognized as {Number, Loca-

[1] http://www.isi.edu/licensed-sw/spade/

tion, Misc, Organization, Person} using the Stanford CoreNLP parser and tagger[2]. The Date and other numeric categories were included within the Number category. These tags weren't always completely orthogonal to each other and sometimes, though not often, the same lexical string appeared in different NE classes - particularly the Misc class. Also for the triplets, the second and the third elements can only belong to one of the following syntactic or semantic labels: {*subject, object, noun, verb, adjective, adverb, other*}

A thorough description of the base models, optimized using the Variational Bayesian framework can be found in (Das, 2011) and is not repeated here for brevity. For optimizing $\boldsymbol{\alpha}$, we resort to the formulations given in (Blei, 2003). The derivative w.r.t. $\alpha_i$ depends on $\alpha_j$ and thus we can resort to Newton's iterative method to find out the maximal $\boldsymbol{\alpha}$ using the gradient and Hessian vector and matrix respectively as in (Blei, 2003). $\Psi'$ is the trigamma function.

$$\frac{\partial \mathcal{L}_{[\alpha]}}{\partial \alpha_i} = M \left( -\Psi(\alpha_i) + \Psi(\sum_{j=1}^{K} \alpha_j) \right)$$

$$+ \sum_{d=1}^{M} (\Psi(\gamma_{d,i}) - \Psi(\sum_{j=1}^{K} \gamma_{d,j})))$$

$$\frac{\partial \mathcal{L}_{[\alpha]}}{\partial \alpha_i \alpha_j} = \partial(i,j) M \left( \Psi'(\sum_{j=1}^{K} \alpha_j) - \Psi'(\alpha_i) \right)$$

$\gamma_d$ is the parameter of the imposed variational Dirichlet distribution over $\boldsymbol{\theta}_d$ as in (Blei, 2003). $\mathcal{L}_{[\alpha]}$ is the topic model objective function, that is used in (Zhu, 2006) for TagLDA and in (Das, 2011) for the Tag$^2$LDA models, but constrained to the terms containing the Dirichlet parameters over $\boldsymbol{\theta}$ only. When $\boldsymbol{\alpha}$ is asymmetric, we optimize each component of $\boldsymbol{\alpha}$ independently. Newton Raphson fixed point iteration based algorithms are used to optimize $\boldsymbol{\alpha}$.

**Differences between MPTag$^2$LDA and corrMPTag$^2$ LDA:** corrMPTag$^2$LDA is a strongly constrained model – A topic's influence over a textual word is obtained by marginalizing out the influences of the corresponding data on it in the document. The more the corresponding data (DL tags) in a document is about a topic the more likely it is that the textual data in the document is also about that topic. This assumption is majorly relaxed in the MPTag$^2$LDA model. In MPTag$^2$LDA, the relation between the DL tags (i.e. $w_n$s) and the textual data (i.e. $w_m$s) are somewhat loose - overall it is possible that two different topics in a document can independently be responsible for the pattern of co-occurrence of the (word, DLtag) ensembles. The TagLDA model(Zhu, 2006) on the other hand does not consider any DL perspective at all and is the least complex of the three classes of the tag-topic models. LDA(Blei, 2003) was not chosen in our experiments due to the consideration of Named Entity classes and the superior performance of TagLDA over LDA in terms of perplexity.

### 2.1 Data Preparation for the Topic Models

While training, each document as a whole, with automatically annotated DL and WL perspectives, was considered without any docset relevancy distinction. While performing inference, we treat each sentence in context. The context is for DL perspective annotation only. A sentence context consists of a central sentence and either its immediate preceding sentence or its immediate succeeding sentence or both depending on the location of the central sentence in the document. For each such central sentence and its context we associate the DL and WL perspectives using the word level annotation vocabulary and the document level tag vocabulary obtained during training. These contexts can be created immediately after a document has been processed for input to the train the models or at a later time for new target documents. Each central sentence during inference was chosen only when at least an automatically detected entity was found. Summary sentences for a docset topic were chosen from among the central sentences collected this way. All stopwords were removed and all words or NE phrases appearing only once across the corpus were removed also.

## 3 The Local Models

In this section we briefly describe the very simple but extremely effective models local to each target docset and each sentence in the target docset. The most time-consuming part for computing the local models was the natural language parsing of sentences. However, all of these computations were done offline as the documents were processed. Five

[2] `http://nlp.stanford.edu/software/corenlp.shtml`

simple local models were considered for each docset and sentences thereof to understand the discriminatory power of the feature sets w.r.t. 5-fold cross-validation accuracies of event category classification of the newswire documents:
1) Collection of the top 20 words (**Bag-tfidf**) across a docset using the $tf \times idf$ weights, where $idf$ has been calculated across the corpus.
2) Collection of the top 20 nouns (**Bag-nn**) which are not proper nouns using $tf \times isf$ weights. The $isf$ (inverse sentence frequency) was calculated only for sentences in the documents within the respective docset.
3) Collection of a bag of 5 most frequent verbs (**Bag-vb**) across all documents in a docset only using $tf \times isf$ weights.
4) Collection of the top 20 nouns (**Bag-nn+vb**) which are not proper nouns and top 5 verbs using $tf \times isf$ weights. The $isf$ (inverse sentence frequency) was calculated only for documents within the respective docset.
5) Collection of top 10 words (**Bag-docfreq**) per document in terms of frequency only.
The cut-off on the frequency counts were tuned through manual inspection of the TAC2010A development set. A full analysis on the implications arising out of an exhaustive enumeration of these numbers was not performed.

The sixth local model that was computed for each sentence, independent of any docset structure, was the set of RS-trees(Soricut, 2003). We followed the work in (Marcu, 1999) to score each node of the RS-trees using the propagation of the salient text spans upward to the root of the tree. The spans i.e. the leaves which were promoted up the RS-tree through internal nodes received a score proportional to maximal heights of such nodes in the tree that contained the promoted spans. In fig. 1b, span 1 only gets a score of 1 while spans 2 and 3 get scores of 3. These scores where then max normalized.

## 4 Summarization Experiments

### 4.1 Summarization Algorithms

In all our experiments we order the sentences (or RST spans) in the descending order of the weights assigned to them. All scores from individual models were normalized between [0,1].

*Sentence scoring strategies based on the topic models*: 1) The sentence weights used were the values of likelihoods from the TagLDA, MPTag$^2$LDA, corrMPTag$^2$LDA models in section 2 corresponding to each sentence fitted to the trained models (c.f. fig. 8a). 2) the weights were obtained by using the expression $\sum_{w_{q,m}} \log p(w_{q,m}|z_{q,m}, \boldsymbol{\theta}_s)$ where $w_{q,m}$ is a word in the query title (average 2-3 words) that is also in the indexed vocabulary, $V$; $s$ is the current sentence whose perspective annotation depends on context (c.f. fig. 8b). 3) Same as (2) except that the summation is over all words in the sentence $s$ that are also in $V$ (c.f. fig. 8c). The first type of weighting has a purely probabilistic interpretation, but the second and third follows from (Nenkova, 2006) and is less intuitive probabilistically.

*Sentence scoring strategies based on local models*: For the non-RS-tree based models, if full sentences were used then they were simply scored by the treating the bag as a list and then using the cosine similarity metric. When RS-tree spans were used to generate sentences, we used the RS-tree span selection criteria only to generate bulleted lists.

*RS-tree span selection criteria*: For every span in the RS-tree, its cosine similarity was calculated using **Bag-tfidf** and query terms. The cut-off for all cosine scores in RS-tree span similarity were set to 0.15 for **nuclei** and 0.1 for selected **satellites** (tuned through summarization performance on TAC2010A dataset). To construct a sentence using RS-tree spans this cut-off criteria was used for both TAC 2010A and 2011A datasets. An appropriate satellite was included if it was the first among the spans or if the number of nuclei found is $\leq 2$.

*RS-tree scoring criteria*: The main score of a span used was the number of leaf nodes in the RS-tree multiplied by the score of the leaf span as mentioned in section 3. This score was also added to the score cosine scores obtained in *RS-tree span selection criteria* but respecting the cut-offs.

Redundancy was handled by adding new sentences or RST spans that did not share 80% of the unigrams or/and bigrams in the set summary sentences previously added. We tried out the MMR strategy of ordering sentences as $s\_score$ = similarity(q, $s_i$) - redundancy($s_i$, $s_j$) $\forall \{i, j\} \in \{s\_in\_docset\}$, with $q$ as the query and $s$ being a sentence, but it proved worse since the similarity and redundancy scores were not homogeneous. In all of our experiments we eliminated all sentences from the summary having $\leq 10$ tokens and $\geq 30$ tokens. Subjective sentence that started with a pronoun enclosed in quotes were not considered. Sentence that had more than 4 numbers - suggestive of a table row or a list of results were eliminated too. These heuristics applied to the sentences created from spans as well.

### 4.2 Evaluation Settings

Due to the lack of resources for manual evaluation, we only used ROUGE(Lin, 2003) as the automatic summary evaluation toolkit. In this paper we report only the ROUGE SU4 scores. The ROUGE S1 and S2 scores are highly correlated to SU4. ROUGE uses a Wilcoxon test to establish confidence intervals. The tools that we used are the RST parser implementation that was used in (Soricut, 2003), the Stanford CoreNLP toolkit and in-house implementations of the extensions to the tag-topic models in (Das, 2011).

Apart from the simple local models acting as baselines, two other baselines were chosen. The Baseline-naive simply returns all the leading sentences (up to 100 words) in the most recent documents. The Centroid(Radev, 2000) baseline is output of MEAD automatic summarizer[3]. A very competitive peer system named CLASSY(Conroy, 2010) was also chosen for comparison. Over the years at the TAC summarization competitions, CLASSY had been fine tuned based on training data from previous year's. For TAC 2011A dataset, it also used very finely crafted vocabulary reflecting the categorical aspects of the information needs. The TopicMarks baseline was obtained from a recent commercial summarization service[4]. Topicmarks summarizes multiple documents by treating all documents as one large document. It does not need any queries and tries to generate key concepts as queries.

### 4.3 Results

In this section we compare and analyze the performances of the different models in the light of perplexity, event category classification cross-validation accuracies and multi-document summarization scores. unless otherwise mentioned, in all figures the legends are read from left to right and from top to bottom corresponding to the groups of bars. In the legends of the figs. 10 and 11, the suffix "FS" means that the summaries were extracted using full sentences and the "RST" suffix means that the sentences were constructed out of salient RS-tree spans.

Figure 3 shows the predictive power of the simple local features for document event categorization during cross- validation. As expected the top 5 docset verbs did not have very high discriminatory power but were not bad either. On the other hand, **Bag-nn+vb** performed consistently high for both datasets. **Bag-tfidf** also performed remarkably well - which is mostly due to the fact that in many cases, it included words from **Bag-nn+vb** as well. **Bag-tfidf** was not restricted to selecting non-Proper nouns only. The performance of **Bag-docfreq** was also encouraging as it did not depend on any docset structure at the time of collection but restricted to a docset as a feature set, shows good generalization power for event classification. The cross-validation graphs in fig. 3 were obtained using the LibSVM library[5] for classification using Support Vector Machines. The predictive performance of these simple features was also intuitive since based on our working hypothesis: frequent nominal mentions and verbs together gives a reader an idea of the "aboutness" of the events unfolding in the relevant documents - albeit in a bag of words form.

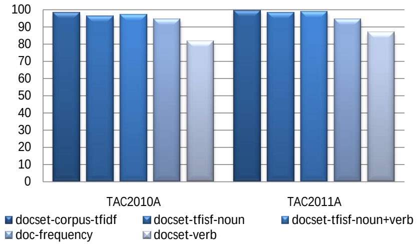


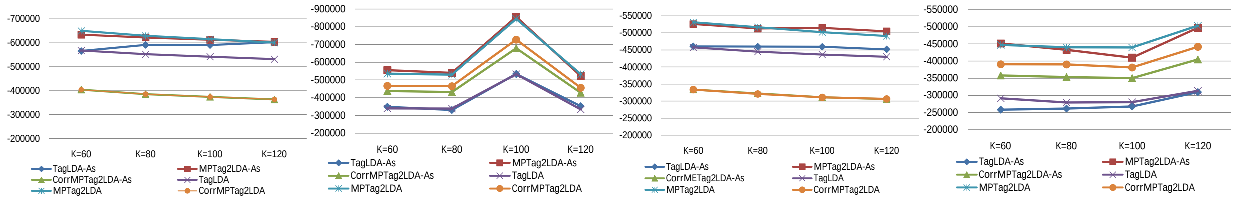
Figure 3: 5-fold Cross-validation accuracies of the local models (bags-of-key_terms) on event category classification of the TAC 2010A/2011A documents. The legend is read from left to right and from top to bottom corresponding to the bar groups for each of the datasets.

From the perplexity point of view, our extended models with the asymmetric Dirichlet prior over the document level topic proportions performs better than the corresponding symmetric case for each class of tag-topic models. The different types of tag-topic models that we considered are TagLDA-As, MPTag$^2$LDA-As and corrMPTag$^2$LDA-As where

[3] http://www.summarization.com/mead/

[4] www.topicmarks.com

[5] http://www.csie.ntu.edu.tw/~cjlin/libsvm/



(a) TAC10A - Training (b) TAC10A - Sentence Contexts (c) TAC11A - Training (d) TAC11A - Sentence Contexts

Figure 4: Evidence Lower BOunds (ELBO)s of the tag-topic models on the TAC 1010A and 2011A datasets in descending order: Lower is better. (Best viewed in color and magnification)

"As (Asymmetric)" means that the components of $\boldsymbol{\alpha}$ in the models in figure 2 can have different values as governed by the co-occurrence pattern of the observations in the data.

From figure 4 we observe that although the correspondence models show the least perplexity on the training set, the TagLDA class of models show better predictive perplexity to fitting the very short sentence contexts as described in subsection 2.1. This reflects the choice of the DL perspective on the assumptions of the model. In our experiments, we artificially created the DL perspective out of frequent words and coherence markers to reflect the approximate attentional state that persists immediately after reading a document. At the document level this indeed lowers perplexity but at the sentence level, due to much lesser context to correspond to at the word level, TagLDA performs better by eliminating the need for correspondence at all. Although the MPTag$^2$LDA performs the worst in terms of perplexity to sentence context fitting, we will see that the trend does not hold true for summarization performance. The models with asymmetric $\boldsymbol{\alpha}$ also show lower perplexity than their symmetric counterparts.

Contrary to perplexity performance on training set, the correspondence class of tag-topic models show very poor generalization power during 5-fold cross-validation for document event classification. We believe that the nature of the DL tagset and the strong constraints on correspondence led to the poor performance. However, it was also interesting to observe that the TagLDA and MPTag$^2$LDA employing the symmetric priors also show similar poor accuracies. MPTag$^2$LDA is loosely constrained on the DL perspective and the TagLDA does not consider that perspective at all. The best performance comes for the latter two class of models but employing an asymmetric prior over the topic proportions with TagLDA-As performing slightly better than MPTag$^2$LDA-As and comparable to that achieved by **Bag-noun+verb** for 120 topics. The event categorization cross-validation graphs for the TAC2010A dataset is is similar to that for the TAC2011A dataset and is shown in fig. 5. Note that the features used here were the $\gamma_{d,i} - \alpha_{d,i}$ for each document $d$ and each topic $i$.

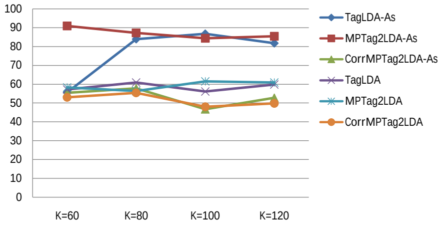


Figure 5: Event classification 5-fold cross-validation accuracies on TAC2010A dataset using tag-topic model features for different number of topics {60, 80, 100, 120} - Higher is better.

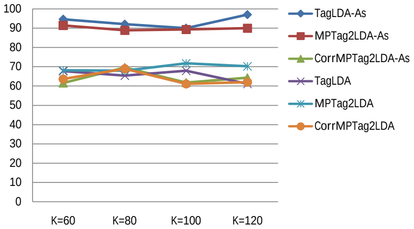


Figure 6: Event classification 5-fold cross-validation accuracies on TAC2011A dataset using tag-topic model features for different number of topics {60, 80, 100, 120} - Higher is better.

This intuitively makes sense if we observe the topics from TagLDA-As from table 1. Table 1 shows some sample (hand selected) latent topics (the marginal $\boldsymbol{\beta}$ from figs. 2a) learnt from the TAC2011A data corresponding to the event categories. The first row shows the conditioning of Topic 30 by the Named Entities and Normal words that receive prominent focus in the topic. The next four rows just shows the marginal topics only. The last row showing topic 32 highlights the use of the asymmetric prior.

| $k$ | Normal_Word | Person | Organization | Location | Misc | Number |
|---|---|---|---|---|---|---|
| 30 | food company recall product dog pet safety die sell death number cat owner test kidney failure | Henderson Sundlof Sarah_Tuite Iams Tuite Nelson Paul_Henderson Burnton | Menu_Foods FDA Iams Food_and_Drug_Administration Wal-Mart Safeway Kroger PetSmart | China United_States Mexico U.S. Canada chinese Arizona Ontario Beijing Shanghai | Menu_Foods cat chinese pet Tootsie canadian north_american bernese room | Monday Friday Saturday Wednesday Tuesday Sunday Thursday March_6 |
| | $\beta_{30}$ : food company pet recall dog cat product kidney Menu_Foods sell safety brand failure test eat death die supplier wheat poisoning | | | | | |
| 7 | $\beta_7$ : flood country Bangladesh river district northern water kill situation relief level government monsoon inundate rain northeastern | | | | | |
| 10 | $\beta_{10}$ : airport police attack car security incident level terminal building London close raise Glasgow british alert arrest Glasgow_Airport | | | | | |
| 75 | $\beta_{75}$ : turtle endanger poach sea fisherman water egg species police jail group beach marine dead Sabah catch fishing fine protect Malaysia | | | | | |
| 0 | $\beta_0$ : Madoff investor money firm fund pay foundation invest son SEC charge jewish New_York hedge business lose part electronic | | | | | |
| 32 | $\beta_{32}$ : the_Associated_Press Timberly_Ross Colo. Coast_Guard European_Union kill Calif. WFP Monday Tuesday government accord country | | | | | |

Table 1: Latent topics from the TagLDA for the TAC2011A data for $K = 80$

Topic 32 (and many other similar topics) have clustered words, like "the_ Associated_Press, kill", that occurred frequently in many documents and were not removed by stopword and low frequency word removal. Words like these do have tendency to distort the latent topic spaces. The asymmetric $\boldsymbol{\alpha}$ prior works by clustering frequently occurring topic dominating words in a few topics and leaving the other topics for data discrimination. This phenomenon occurred much less with Tag$^2$LDA class of models because of the correspondences to the particular nature of the DL tags we considered and the intrusive topic dominating words were assigned low probabilities. With the power of superior event discrimination from the asymmetric versions of the tag-topic models, we used only these for multi-document summarization experiments.

Figure 8 shows the ROUGE SU4 scores of the summaries obtained by weighting words from full sentences obtained from the different tag-topic models for the TAC2011A dataset. The results for the TAC2010A dataset as they are very similar and are shown in fig. 7. The sentence in context likelihoods gave best results for MPTag$^2$LDA achieving almost the same scores for 100 topics for the weighting with query terms only. For the second kind of weighting also, MPTag$^2$LDA performs as good as TagLDA at higher $K$s due to simultaneous topical agreements with two perspectives. In the third type of weighting, query drifting is responsible for the lowering of the scores. The correspondence tag-topic model was not performing well mostly because the correspondences between the DL perspective and the words were shared across different event categories - for e.g., the DL tags containing lemmatized words like "die" is easily associated with various entities and concepts across "Attacks," "Natural Disasters" and "Heath" categories. However, all of them performed statistically as well as the robust algebraic Centroid algorithm. The number of topics did not show much discrimination on the ROUGE SU4 scores.

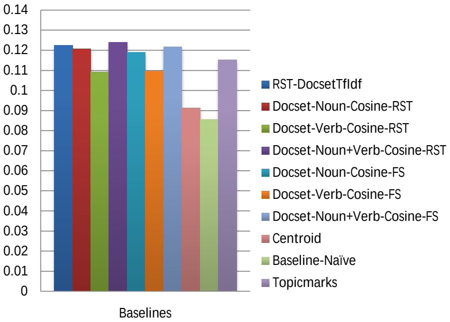


Figure 9: ROUGE SU4 scores on TAC2010A dataset for Local models - Higher is better.

Figures 9 and 10 show the power of simple local (mostly vector-space) models on the summarization performance over feature sets that show high event categorization power. Performances on

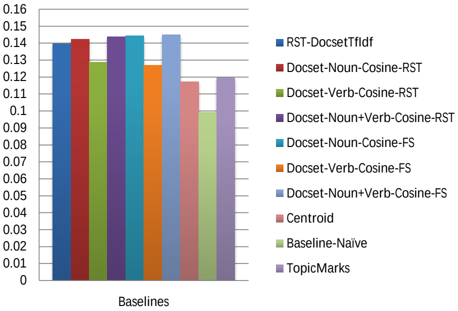
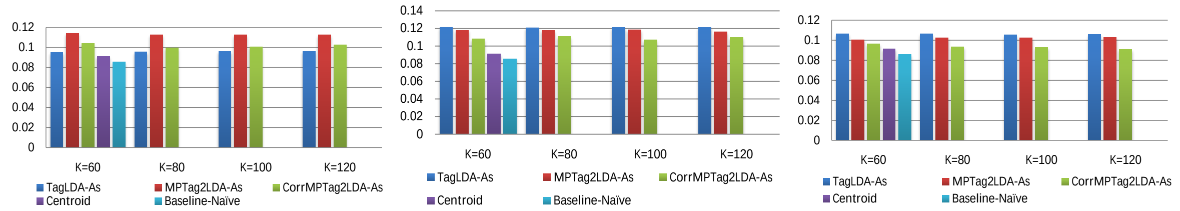


Figure 10: ROUGE SU4 scores on TAC2011A dataset for Local models - Higher is better.

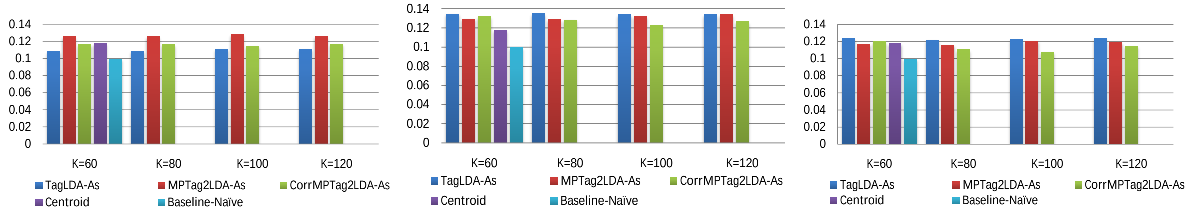


(a) Sentence likelihoods

(b) Cumulative mass of query words in the sentence

(c) Cumulative mass of all indexed words in the sentence

Figure 7: ROUGE SU4 scores for summaries obtained from purely from tag-topic models on the TAC 2010A dataset: Higher is better. (Best viewed in color and magnification)



(a) Sentence likelihoods

(b) Cumulative mass of query words in the sentence

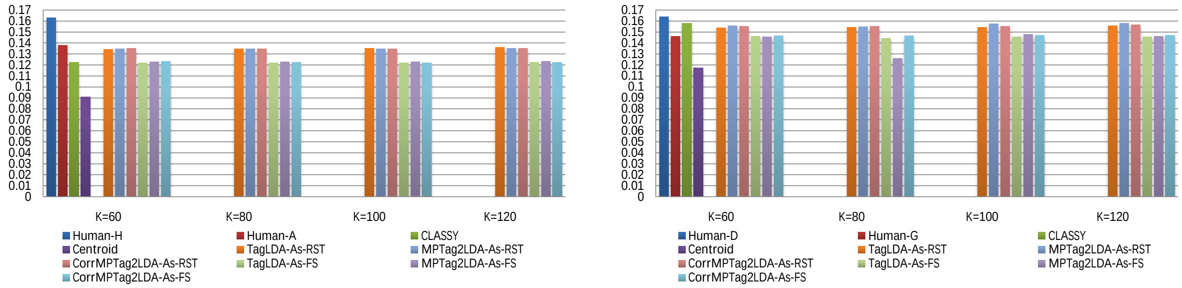
(c) Cumulative mass of all indexed words in the sentence

Figure 8: ROUGE SU4 scores for summaries obtained from purely from tag-topic models on the TAC 2011A dataset: Higher is better. (Best viewed in color and magnification)

the TAC2010A dataset is shown in fig. 9 and that on the TAC2011A dataset is shown in fig. 10.

The summarization performance of the simple local models show that in absence of a query, **Bag-noun+verb** represents the information need quite well provided there is a docset of relevant documents. It was very interesting to see the effect of document relevancy on multi-document summarization based just on key verbs representing events. The RST version of that is significantly better than Centroid at 95% confidence. Even the RST- DocsetTfIdf also performed very well using the *RS-tree scoring criteria*.

Figure 11 shows the ROUGE SU4 scores of the summaries from TAC 2010A and 2011A datasets using a combination of local and global tag-topic models. We observe a major boost in ROUGE scores when a simple but intuitive additive fusion function was employed. We added the sentence likelihoods and the cosines from the local vector space model using **Bag-noun+verb** features. Even summaries with full sentences for TAC2010A were as good as the CLASSY system(Conroy, 2010). With the bulleted list i.e. RS-tree span summaries our system beats CLASSY on the TAC2010A dataset at 95% confidence. In TAC2011, the CLASSY system was modified to use bigrams and more categorical aspect specific vocabulary. Our system did not use any hand crafted vocabulary for aspect matching and is based on the intuitions of a reader's behavior. Both CLASSY and our system performed equally well and actually crossed the median human score on TAC2011A ROUGE SU4 evaluation (only the lowest the highest are shown in fig. 11). The systems also beat the baseline models at 95% confidence interval. A possible explanation for crossing median human scores is that the assessors were free to choose any vocabulary that had seemed fit rather than using those occurring only in the input documents. In our experiments we found that the cumulative probability mass weighting did not help in the fusion of sentence scores from the local models. It is possible that such fusions were competitive rather than collaborative. A language independent version of our system can be built using **Bag-tfidf** and using positional information at word level (or other markup tags) and (optionally) a DL perspective as well. The inequalities of the likelihood contributions of the tag-topic models (c.f fig. 8a) was compensated by the relative ordering. However, given a choice on multi-document summarization, it is better to use TagLDA-As or MPTag$^2$LDA-As based on



(a) TAC10A - Human summaries, CLASSY, Centroid and Proposed approaches for different number of topics

(b) TAC11A - Human summaries, CLASSY, Centroid and Proposed approaches for different number of topics

Figure 11: ROUGE SU4 scores on TAC 2010A and 2011A datasets - Higher is better. (Best viewed in color and magnification)

this type of DL perspective. The event classification power of the tag-topic models was a major indicator of their summarization power when considering sentence likelihoods alone.

The use of RS-tree spans using a thresholding criteria allows us to use the spans as bulleted lists and pack more information in less words. An example summary using RS-tree spans is shown in table 2 for the "sleep deprivation" topic in TAC 2011A. The summaries in 2 actually had low ROUGE scores since it was much harder to assess which fact should really be used in the summary. The choice of successive salient RS-tree spans also needs to be bettered for readability purposes - the satellite "who slept for the recommended seven hours or more a night. than" was not included in the first bullet as it failed the filtering criteria. We also noticed that varying the number of topics simply permuted the order of sentences.

## 5 Previous Experiments

In this section we highlight the key steps that we used in our original submissions to the Guided and Multilingual Summarization tasks. In our initial design, this step involved the local docset topic structure (i.e. the local set of relevant documents) for weighting some terms. The main motivation to use the docset topic structure was to incorporate as much information into the background models as possible. Note that in all the term vocabularies created for datasets in English or otherwise, stop words were identified and eliminated. This was done so that firstly the statistical topics don't get dominated by stopwords and secondly, we don't score sentences over stopwords. As we have seen in earlier sections, the problem with stopwords in topic modeling can be addressed with an asymmetric topic proportion prior. In the following subsections, we briefly go over our **initial system design** and show the official scores as a result of the shortcomings.

### 5.1 Creating DL Perspective for Guided Summarization Task

The following list provides a gist of the steps which were followed to process each docset initially.

- Collect all terms in a docset with their parts of speech.
- Order terms by score following the usual $tf - idf$ convention but $idf$ restricted only to the docset. The score for a term is $\sum_{t \in docset\_topic} tf_{t,d} \times idf_{t,docset\_topic}$ computed over all $d \in docset\_topic$. The main intuition behind this scoring was to find terms that had a balance between being mentioned a number of times in some documents but not all in the documents of the docset topic.
- Collect all terms from the docset with verb as their part-of-speech and their counts.
- Collect all triplets i.e. coherence markers as in section 2. However all triplets that contain verbs as both the second and third elements were not considered. This was done since important verbs and their dependents are considered in the WL annotation.

We then created the DL tagset by iterating over each document in the docset through the following steps:

| **Model Summary** | **Summary from MPTag$^2$LDA-As (K=100) using bulleted lists** | **Summary from MPTag$^2$LDA-As (K=100) using full sentences** |
|---|---|---|
| [Research has found that sleep deprivation is associated with serious health problems such as depression, obesity, cardiovascular disease and diabetes.] [Lack of sleep adversely affects memory function and athletic performance.] [Sleep disorders are common in people 60 and over.] [Women's health is at much greater risk than men's.] [Sleep-deprived adolescents are more likely to use alcohol and tobacco.] [Sleep-deprived children can exhibit ADHD-like behavioral problems.] [Sleep medications are increasingly prescribed for children, but their safety and effectiveness are unknown.] [For adults, napping has rejuvenating effects and boosts alertness, performance & productivity.] | • who slept less than or equal to 5 hours a night were twice as likely to suffer from hypertension than women<br>• children ages 3 to 5 years get 11-13 hours of sleep per night ;<br>• A previous study on sleep deprivation - less sleep resulted in impaired glucose metabolism,(*)<br>• sleeping less than eight hours at night, frequent nightmares and difficulty initiating sleep were significantly associated with drinking.<br>• athletes often aren't counseled on the value of adequate sleep, to sleep deprivation,(*)<br>• A single night of sleep deprivation can limit the consolidation of memory the next day, | [The study found that a sleep deficit built up over just five nights can significantly impair heart function.] [The study, published in a recent issue of the American Journal of Geriatric Psychiatry, spoke about the high rate of sleep complaints among the patients.] [Researchers found no difference between men sleeping less than 5 hours and those sleeping 7 hours or more.] [The results showed that sleeping less than eight hours at night, frequent nightmares and difficulty initiating sleep were significantly associated with drinking.] [Older folks nap, partly because they don't sleep as soundly at night.] [Sleep is critical to preparing the brain to lay down memories the next day.] |

Table 2: 100-word Summaries for the harder information need on "Sleeping Deprivation" in TAC2011 dataset. Individual sentences are square bracketed. A (*) indicates that the bullets came from the same document.

1. Processed non-stopwords only
2. Collected around top 20 terms that rely on cumulative docset $tf_{t,d} \times idf_{t,docset\_topic}$ scoring (per docset)
3. Collected triplets that do not contain both first and second elements as verbs for e.g. attacking→attack is fine because of the lemmatization (per document)
4. Collected terms that matched the top 20 $tf_{t,d} \times idf_{t,docset\_topic}$ terms from a docset. (per document)
5. Collected top 5 most frequent terms (per document). This was done only if there were no triplets or there were no words that matched the top 20 docset specific important words. This was done only for including short irrelevant documents into the training index.

### 5.2 Creating DL Perspective for Multilingual Summarization Task

The process of obtaining the document level perspective for the multilingual documents is very similar to that for the Guided Summarization task documents. The steps are listed as follows:

- Sentence splitting was done using standard end sentence markers and the unicode sentence delimiter character in Hindi documents.
- Collected stopwords by translating the stopwords in English to the respective languages. The translation was done using Google Translate API service. Although we could have eliminated stopwords by looking at the corpus wide (not docset wide) cumulative tf-idf statistics for multilingual tokens, we chose not to do it.
- Collected terms (without any form of lemmatization but not stopwords) in a docset. We do not use any part-of-speech tagger and hence have no part of speech information. This is done for even the English documents in the multilingual task.
- Ordered terms by score following the usual tf-idf convention but idf again restricted only to the docset. The score for a term is again $\sum_{t \in docset\_topic} tf_{t,d} \times idf_{t,docset\_topic}$ computed over all $d \in docset\_topic$.
- Collected all terms that had been mentioned in the contextual sentences without regard to their part-of-speeches or dependency labels. So for example, the word "earthquake" will be represented as "(earthquake, xx, xx)". where "xx" is a place-holder for an unknown part-of-speech or dependency label. So if "earthquake" appears in the list of the docset specific top tf-idf words, then the DL perspective will have both the tags represented as "(earthquake, xx, xx)" and "earthquake".

We then created the DL tagset by iterating over each document in the docset in the following steps:

1. Processed non-stopwords only

2. Collected around top 20 terms that rely on cumulative docset $tf_{t,d} \times idf_{t,docset\_topic}$ scoring (per docset)

3. Collected cross-sentence triplets but without any form of lemmatization or stemming - for the English documents also.

4. If a word matched the top 20 $tf_{t,d} \times idf_{t,docset\_topic}$ (per document) the term was used as a DL perspective tag

5. Collected top 5 most frequent terms (per document). This was collected only if there were no triplets or there were no words that matched the top 20 docset specific important words. This was again done only for including short irrelevant documents into the training index.

We next focus on creating the word level annotations to be used in the word level perspectives.

### 5.3 Creating WL Perspective for Guided Summarization Task

The listed guidelines were followed for annotating each word in every sentence. Note that in the guided summarization task, named entities are treated as phrases.

- We considered a set important dependency relations {Nominal_Subject, Passive_Nominal_Subject, Controlling_Subject, Direct_Object, Indirect_Object, Object_Of_A_Preposition, Relative} as obtained by running the Stanford CoreNLP toolkit.

- Out of the top 10 non-stop word verbs in every docset, we identified all children of those verbs following the important dependency relations. We also identified all parents of such verbs following the important dependency relations. The verbs together with governor parents and dependent children were identified as "*subjective*" words. Intuitively, the most frequent verbs (which are not stop-words) collected from a particular docset usually identifies the "what is happening" attribute of the category of the document set. This is particularly true of multiple news reports being gathered into a pre-determined topic cluster.

- All Named Entities were automatically recognized as {Date, Location, Misc, Organization, Person} using the Stanford CoreNLP tagger.

- Words other than entities were annotated as *Normal_Words*. However, the *subjective* annotation overrides all other WL annotations.

### 5.4 Creating WL Perspective for Multilingual Summarization Task

Due to deliberately not keeping any syntactic or semantic (e.g. named entity) annotation for the multilingual documents, we just annotated every word in every document with a choice of 5 possible positions as WL annotation categories. We segregated every document into 5 positional zones - {begin, begin→middle, middle, middle→end, end} and the words falling into those zones were annotated as such. The intuitive idea is that if different zones generate the same word, then that word is identified more by its latent topical information rather than the positional zones and vice-versa if otherwise.

### 5.5 Data Preparation for the Topic Models

The data preparation for our initial submissions was done in the same way as mentioned in subsection 2.1. For the Guided Summarization task, each central sentence during inference was chosen only when at least an automatically detected entity was found. For the Multilingual task, every sentence of length at least 10 words was considered. We created a collection of such sentence contexts for every docset topic. Summary sentences for a docset topic were chosen from among the central sentences collected this way.

### 5.6 The Local Models

Here we briefly describe the models local to each docset and each sentence in the docset. Note that, in the recent modifications, we have vastly improved upon the local models by using Rhetorical Structure trees that directly addresses the aspects of the information needs not covered by Named Entity categories as well as data compaction. This is the part that had not been properly optimized in our initial experiments. Three local models were considered for each docset and the sentences thereof to collect the following information:

a) Collection of a bag of all nouns (**Bag-nn**) which are not proper nouns from the DL perspective (see subsection 5.1)

b) Collection of a bag of all verbs (**Bag-vb**) which are not stopwords from the WL perspective (see subsection 5.3)

c) Collection of the dependency parsing outputs for each sentence in the docset. The output was produced using Stanford's CoreNLP parser.

### 5.7 Query Dependent Scoring in the Guided Summarization Task

Given a short query title, the following steps were followed to accumulate scores from the global and local models.

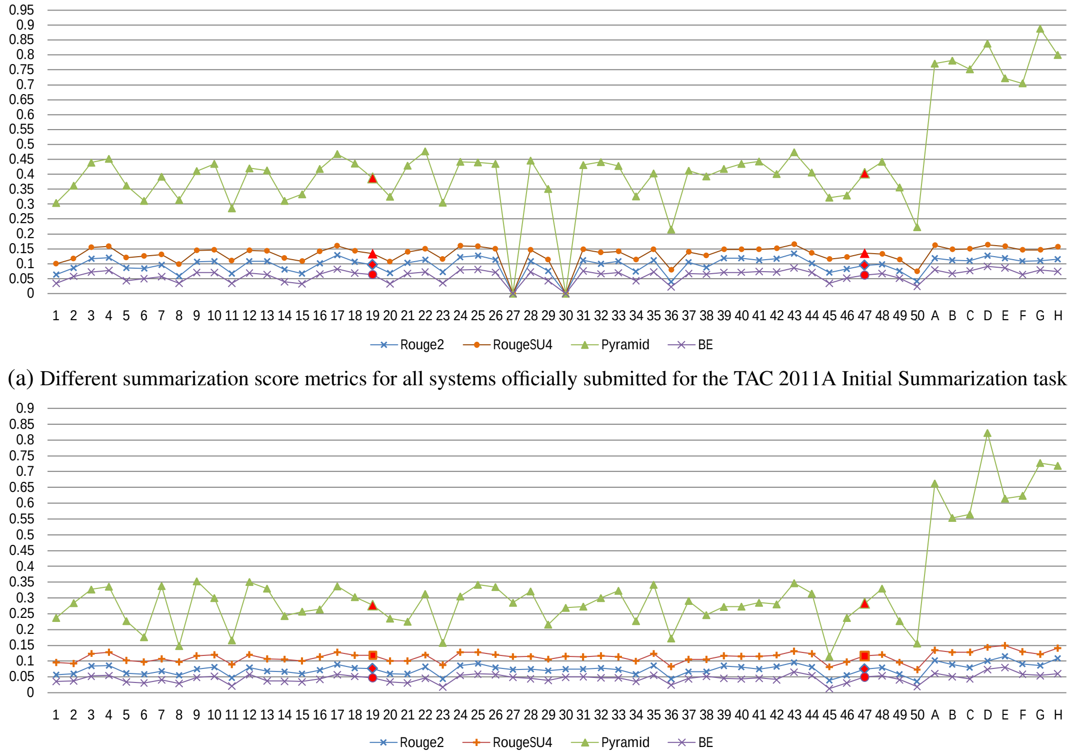
1. Sum of probability masses of the query words **only** in the sentence. This is exactly same as that mentioned in subsection 4.1. This score is obtained using the background/global models. This type of scoring is similar in spirit to that used in Vanderwende et. al. (Vanderwende, 2007).
2. Cosine similarities of the query with sentences w.r.t local sentential term frequencies.
3. Number of important verbs (see section 5.3) in the sentence.
4. Dependency scoring for every sentence in a docset w.r.t. every query term and sentence term pair:
   - Find the paths following dependency relations from source (query term) to target (sentence term). The source and the target are interchanged if no paths are found.
   - If there is one or more paths, then we do the following: i) for each path calculate the number of times words in the path, say of length $L$, match Bag-nn, say $\mathcal{N}$, and also calculate the number of times words in the path match Bag-vb, say $\mathcal{V}$. ii) For each path, we then update a score for the sentence as $(\mathcal{N}/L_{cumulative}) \times \mathcal{V}$, where $L_{cumulative}$ is the sum of the path lengths $L$ if there are more than one path found in the dependency graph for the query and sentence term pair. This is something which has not been tuned at all and has been a primary cause in the failure of the local model to provide sufficient boost over the global model.

**The dependency model was rejected in our new version in favor of sentential RS-trees**.

All central sentences in a docset topic were scored in these ways from different models and an Maximum Marginal Relevance (MMR)(Carbonell, 1998) based greedy algorithm was chosen to order the sentences using the sentence-sentence similarity matrices. The objective function to maximize was $score_{global+local\,models}(query, sentence_i) - redundancy(sentence_i, sentence_j)\,\forall\,\{i,j\} \in \{sentence\_index_{docset\_topic}\}$

For the "B" timeline, a cosine similarity between the current sentence and the previous summary was also factored into the MMR formulation. Also for the "B" timeline summaries, the score w.r.t. the query and the sentence did not involve any dependency scoring or important verb coverage scoring. This was done as part of tuning on the TAC2010 development set. Although MMR was employed, since the scoring function involved real numbers as outputs from several models, we added the check that if any two sentences were 50% common in bigrams or unigrams (with bigram checking coming first) then they were considered as duplicates. These needed to be done since the scores were not normalized and also the similarity to query and redundancy between sentences in the MMR formulation were not consistent.

Figure 12 shows the results of all the systems that were submitted as official runs in the TAC 2011 guided summarization task competition. The red markers in both figs. 12a and 12b, show the scores obtained by our initial submissions (run ids – 19 and 47) on different scoring criteria. The two runs merely differed by query expansion based on the local docset bags of nouns. Our initial scores were statistically much less significant than that for the human summaries as well as CLASSY(Conroy, 2010)(run id – 25) under all evaluation metrics. The update task was minimally addressed and thus mediocre scores were expected. Fig. 11, on the other hand, shows that using our new modifications, we have significantly outperformed our own official submission for the Guided Summarization task on initial summary generation.



(a) Different summarization score metrics for all systems officially submitted for the TAC 2011A Initial Summarization task

(b) Different summarization score metrics for all systems officially submitted for the TAC 2011B Update Summarization task

Figure 12: Different summarization score metrics for all systems officially submitted to the Guided Summarization task for TAC 2011

### 5.8 Query Independent Scoring in the Multilingual Task

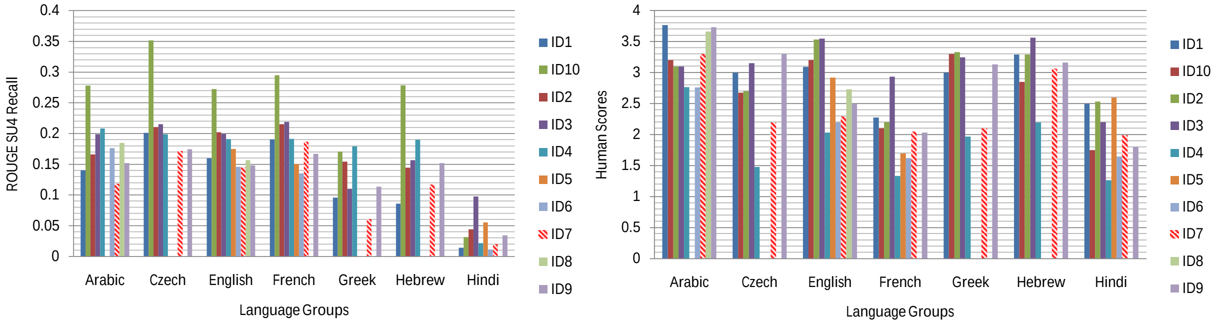
We chose not to generate any query from the docsets whatsoever. Instead, as described in section 5.5, we inferred the likelihoods of the sentence contexts to the corrMPTag$^2$LDA model. The number of latent topics was set to 30 since there were 10 docset topic clusters. A single corrMPTag$^2$LDA model was fit to each language. Since likelihoods tend to be higher for shorter sentences we only ordered those sentences in the descending order of likelihoods which have a cut-off length of 20 words. The summaries generated involved minimum effort and no language specific information other than the use of language specific stopwords. This score does have a probabilistic interpretation - similar to the $Product_{CF}$ in (Nenkova, 2006).

Although ROUGE scores put our system (run id – 7) in the last place, the human scores do not do so. This is in part due to the small number of topic clusters used in the pilot MultiLingual Summarization task where the ROUGE scores do not correlate as well to the human evaluations(Dang, 2006). Based on our new evaluations on the use of topic models for document event discrimination, the use of corrMPTag$^2$LDA was not right in the hindsight. This fact coupled with the usage of only a symmetric Dirichlet topic proportion prior had indeed caused a catastrophic decrease in ROUGE-SU4 scores.

Due to time constraints, we were not able to conduct a thorough set of experiments with our new modifications on the update task as well as the multilingual task. We plan to do that in the future. It will be very interesting to see if local models **from the updated documents only** are able to boost scores as opposed to actually looking at the previous summary to avoid redundancy.

## 6 Conclusion

We have made significant progress in generating better summaries from those that were submitted in the official runs. In summary, we have shown that it is possible to use unsupervised models that do latent structure discovery of (word, annotation) ensembles



(a) Scores based on ROUGE-SU4 for all systems officially submitted for the TAC 2011 Multilingual task

(b) Scores based on manual evaluation for all systems officially submitted for the TAC 2011 Multilingual task

Figure 13: ROUGE-SU4 Recall and Human Evaluation scores for all systems officially submitted to the Multilingual Summarization task for TAC 2011. The failure of using sentence likelihoods from corrMPTag$^2$LDA is apparent.

in text which is resilient to the occurrences of stopwords. At the same time, the likelihoods of the local contexts that are fit to the models together with simple local models that capture relevancy in target documents show impressive multidocument summarization power. The use of RS-trees to generate bulleted list summaries also shows promising result within this framework. As a future work we want to experiment with bigrams or dependency triples as vocabulary units to see if those can further improve the likelihoods from within the exploratory models themselves. To address the issue of readability involving coherence we would like to follow the traveling salesman approach (Conroy, 2010) to order summary sentences using both lexical similarity and coherence indicators(Barzilay, 2005). We also plan to test our method on other genres of text (for e.g. scientific literature).

## References

Lin, Chin-yew and Hovy, Eduard. 2000. *The Automated Acquisition of Topic Signatures for Text Summarization*. In proc. of the COLING conf., 495–501.

Barzilay, Regina and Lapata, Mirella 2005. *Modeling local coherence: an entity-based approach*. In proc. of the 43rd ACL conf., 141–148.

Lin, Chin-Yew and Hovy, Eduard 2003. *Automatic evaluation of summaries using N-gram co-occurrence statistics*. In proc. of the 2003 NAACL HLT conf., 71–78.

Grosz, Barbara J. and Weinstein, Scott and Joshi, Arvind K. 1995. *Centering: A Framework for Modeling the Local Coherence of Discourse*. In Computational Linguistics, volume 21:203–225.

Das, Pradipto and Srihari, Rohini and Fu, Yun 2011. *Simultaneous Joint and Conditional Modeling of Documents Tagged from Two Perspectives*. In Proc. of the 20th ACM CIKM conf..

Vanderwende, Lucy and Suzuki, Hisami and Brockett, Chris and Nenkova, Ani 2007. *Beyond SumBasic: Task-focused summarization with sentence simplification and lexical expansion*. In Information Processing and Management, volume 43:1606–1618.

Nenkova, Ani and Vanderwende, Lucy and McKeown, Kathleen 2006. *A compositional context sensitive multi-document summarizer: exploring the factors that influence summarization*. In proc. of the 29th ACM SIGIR conf., 573–580.

David M. Blei and Andrew Y. Ng and Michael I. Jordan 2003. *Latent Dirichlet Allocation*. Journal of Machine Learning Research, volume 3:993–1022.

Wallach, Hanna M. and Mimno, David and McCallum, Andrew 2009. *Rethinking LDA: Why Priors Matter*. In proc. of NIPS conf.

Dang, Hoa Trang 2006. *DUC 2005: evaluation of question-focused summarization systems*. In proc. of the workshop on Task-Focused Summarization and Question Answering, 48–55.

Radu Soricut and Daniel Marcu 2003. *Sentence Level Discourse Parsing using Syntactic and Lexical Information*. In Proc.of the NAACL HLT conf..

Haghighi, Aria and Vanderwende, Lucy 2009. *Exploring content models for multi-document summarization*. In Proc. of NAACL/HLT, 362–370.

Mann, William C and Thompson, Sandra A 1988. *Rhetorical Structure Theory: Toward a Functional Theory of Text Organization*. Text, volume 8:243–281.

Ryan T. McDonald 2007. *A Study of Global Inference Algorithms in Multi-document Summarization*. In proc. of ECIR, 557–564.

Hal Daumé III and Daniel Marcu 2006. *Bayesian Query-Focused Summarization*. In proc. of the ACL conf..

D. Radev and H. Jing and M. Budzikowska 2000. *Centroid-based summarization of multiple documents: Sentence extraction, utility-based evaluation, and user studies*. In proc. of ANLP/NAACL Workshop on Summarization, 21–29.

Genest, P. and Lapalme, G. and Yousfi-Monod 2010. *HEXTAC: the Creation of a Manual Extractive Run*. In proc. of the NIST Text Analysis conf. (TAC).

Conroy, John M. and Schlesinger, Judith D. and P.A. Rankel and O'Leary, Dianne P. 2010. *Guiding CLASSY Toward More Responsive Summaries*. In proc. of the NIST Text Analysis conf. (TAC).

Marcu, Daniel 1999. *Discourse trees are good indicators of importance in texts*. Advances in Automatic Text Summarization: Editors Inderjeet Mani and Mark Maybury, 123–136.

Marcu, Daniel *The rhetorical parsing of unrestricted texts: a surface-based approach*. Computational Linguistics, volume 26:395–448.

Asli Celikyilmaz, and Dilek Hakkani-Tr 2011. *Discovery of Topically Coherent Sentences for Extractive Summarization*. In proc. of the 49th NAACL/HLT conf. 491–499.

Xiaojin Zhu and David Blei and John Lafferty 2006 *TagLDA: Bringing document structure knowledge into topic models*. In UWisc Technical Report TR-1533.

Jaime Carbonell and Jade Goldstein 1998. *The use of MMR, diversity-based reranking for reordering documents and producing summaries*. In proc. of the SIGIR conf., 335–336.