Proceedings of the Twenty-Sixth AAAI Conference on Artificial Intelligence

Generating Coherent Summaries with Textual Aspects

Renxian Zhang **Wenjie Li** **Dehong Gao**

The Hong Kong Polytechnic University
{csrzhang, cswjli, csdgao}@comp.polyu.edu.hk

Abstract

Initiated by TAC 2010, aspect guided summaries not only address specific user need, but also ameliorate content level coherence by using aspect information. This paper presents a full fledged system composed of three modules: finding sentence level textual aspects, modeling aspect based co herence with an HMM model, and selecting and ordering sentences with aspect information to generate coherent summaries. The evaluation results on the TAC 2011 da tasets show the superiority of aspect guided summaries in terms of both information coverage and textual coherence.

Introduction

Traditionally, text summarization techniques are developed to maximize the coverage of salient information in the original text. Many popular models compute information salience from the distributional frequency of textual units. But if we specify the particular kinds of information to be covered, the frequency-based approach is not guaranteed to work. For example, we require the *cause* of an accident to be included in an extractive summary of a 30-sentence news report, which mentions the target information without using the word *cause* or its synonyms in only 1 sentence. Collecting frequent words and sentences may not help.

Such particular kinds of information are termed **aspects** by the summarization track of TAC 2010 to "encourage a deeper linguistic (semantic) analysis"[1]. Note that "aspects" here are "textual aspects" acting as semantic components, which are different from "verb aspects" (e.g., *simple*, *progressive*, *perfect*) in grammar analysis, "product aspects" (e.g., *price*, *service*, *value*) in opinion mining, etc.

In this work, we are committed to generating aspect-guided summaries, which has greater significance than meeting the TAC agenda.

As an upgrade of query-focused summaries, aspect-guided summaries are more focused on user need, consisting of finer-grade semantic elements. More importantly, aspects enable us to produce **content-level coherent** summaries. Given a set of aspects {*time*, *place*, *casualties*, *cause*, *countermeasures*} for an accident report, we can extract and arrange summary sentences according to the natural order and logical development among aspects such that, for example, *time* and *place* are preferably mentioned together, and *countermeasures* should follow *casualties*. A summary constructed in this way is ideal in that it 1) addresses specific and semantically structured user need, and 2) achieves good coherence on the content level.

To generate aspect-guided summaries, we are faced with two major challenges. 1) How do we find aspects since they are content units hidden beneath the surface? 2) How do we use aspects to model textual coherence in order to generate coherent summaries? Their solutions constitute our contributions in this work. Specifically, we

- develop the novel meta-phrase features to help find aspect-bearing sentences, formulated as a multi-label classification problem;
- model aspect-based coherence with an HMM model, which proves superior to a previous model that does not use aspect information;
- propose a summarization approach that leverages recognized aspects and aspect-based coherence, which performs very competitively on a benchmark dataset.

In the next section, we review previous work that helps to shape up the current endeavor. That is followed by three sections that address three major modules: aspect recognition, aspect-based coherence modeling, and aspect-guided summarization. Then we present experimental results, and finally we conclude the paper with a future direction.

Related Work

The recent interest in aspect-guided summarization is partly inherited from query-focused summarization. Queries are usually handled to extract relevant sentences by Information Retrieval (IR) techniques such as query expansion

[1] http://www.nist.gov/tac/2010/Summarization

(Li et al. 2006; Vanderwende et al. 2007) built into statistical (Daumé and Marcu 2006), graph-based (Wan et al. 2007), and learning-based (Fuentes et al. 2007; Schilder and Kondadadi 2008) models. The state of the art from this camp, however, does not fit the nature of aspects as particular kinds of information that a summary should include.

Li et al. (2011) apply an unsupervised and topic model-based approach to aspect-guided summarization. The major limitations are that each sentence is assumed to have a specific aspect (many TAC document sentences do not) and that the aspects identified by sentence clustering do not necessarily match the actually expected aspects.

In contrast, Information Extraction (IE)-based summarization is a better match for aspects. Teufel and Moens (2002) summarize scientific articles by extracting sentences of certain "rhetorical statuses" – domain-specific aspects. Ji et al. (2011) propose to find facts about entities, events, and relations to generate query-focused summaries. We complement their work with new evidences from aspect-guided news summarization.

We find aspects on the sentence level, a sentence classification problem. Related works focus on domain-specific sentence classes, such as *Background*, *Topic/Aboutness*, etc. for science research papers (Teufel and Moens 1999), *Introduction*, *Method*, etc. for medical abstracts (McKnight and Srinivasan 2003), and *Bio*, *Fact*, etc. for biographies (Zhou et al. 2004). But we are not aware of prior works dedicated to sentence-level news aspect classification.

Recently, coherence in summaries draws much attention, with much effort to model its local (Barzilay and Lapata 2008) and global features (Barzilay and Lee 2004) and their combination (Elsner et al. 2007). Our modeling of aspect-based coherence improves on the framework in (Barzilay and Lee 2004), which models sentences as observed sequences emitted from hidden content topics. A crucial difference between our model and theirs is the use of aspects as an intermediary between sentence and words.

Aspect Recognition

In this section, we explain how aspects are found on the sentence level – appropriate for extractive summarization.

Feature Extraction

Since aspects may be hidden under the literal content, we devise a new type of features: meta-phrase features.

We define a meta-phrase as a 2-tuple (m_1, m_2) where m_i is a word/phrase or **word/phrase category**, which is a **syntactic tag**, a **named entity (NE) type**, or the special /NULL/ tag. Syntactic tags represent the logical and syntactic attributes of words in a sentence, including logical constituents (/PRED/ for predicate, /ARG/ for argument) and grammatical roles (e.g., /dobj/ for direct objet, /nn/ for noun modifier). A predicate can be a verb, noun, or adjective and an argument is a noun. The combination of syntactic tags and/or words gives rise to meta-phrases of the **syntactico-semantic pattern**, including the predicate-argument pattern and the argument-modifier pattern. Table 1 has examples.

NE types represent the semantic attributes of special NPs in a sentence, which are indicative of particular aspects. We use NE types such as person (/PER/) and organization (/ORG/). The combination of NE type and/or NE word/phrase gives rise to meta-phrases of the **name-neighbor pattern**, including the left neighbor-name pattern and the name-right neighbor pattern. Examples are provided in Table 1.

Syntactico semantic patterns	Predicate argument	*linked fen phen* → (/PRED/, /dobj/)
	Argument modifier	*Clinic study* → (/nn/, /ARG/)
Name neighbor patterns	Left neighbor name	*a Mayo Clinic* → ('a', /ORG/)
	Name right neighbor	*Mayo Clinic study* → (/ORG/, 'study')

Table 1: Meta-phrase patterns and examples

In the above, we have only shown one of the extractable meta-phrases from tag/word combinations. For syntactico-semantic patterns, two related words and their syntactic tags give a total of 4 combinations. For example,

linked fen-phen {
(/PRED/, /dobj/)
(/PRED/, 'fen-phen')
('linked', /dobj/)
('linked', 'fen-phen')

For name-neighbor patterns, an NE or its type alone (with the /NULL/ tag) or with its left/right neighbor gives 4 combinations as shown below.

Mayo Clinic study {
(/ORG/, 'study')
(/ORG/, /NULL/)
('Mayo Clinic', 'study')
('Mayo Clinic', /NULL/)

Such syntactico-semantic and name-neighbor meta-phrases are designed to capture syntactic relations and NE contexts at different levels of abstraction.

Name-neighbor meta-phrase extraction relies on NE recognition; syntactico-semantic meta-phrases are extracted in three scans via dependency parsing: 1) Scan for all predicate-argument pairs in the sentence from dependency relations: nominal subject, direct object, agent, etc.; 2) Scan for all nominal argument modifiers from dependency relations: noun modifier, appositional modifier, etc.; 3) Scan for all adjectival argument modifiers from the dependency relation of adjectival modifier.

Multi-label Classification

One sentence may be associated with multiple aspects, so aspect recognition on the sentence level is a multi-label classification problem.

Label combination and binary decompositions (Boutell et al., 2004; Tsoumakas and Katakis, 2007) can be used to transform multi-label classification to single-label classifications. The former maps the original k label sets to the 2^k label power sets by transforming all distinct label subsets into single label representations. The latter transforms the original k-label classification into k single-label classifications before aggregating the k classification results to obtain the final result.

A potential problem with label combination (LC) is that there may not be sufficient training data available for each transformed single-label class, whereas binary decomposition (BD) assumes label independence which does not necessarily hold. In the "Experiments" section, we will show that BD performs better for our task.

Semi-supervised Learning

We observe that for this task, classification accuracy may suffer from insufficient training data and a model learned from limited training data may not adapt well to unseen data. For example, in the TAC data used in our experiments, "health and safety" articles can range from Chinese food safety to protective helmets in the United States.

A promising answer to those issues lies in transductive SVM (Vapnik, 1998; Joachim, 1999), which predicts test labels by using the knowledge about test data. So it addresses both training (labeled) data deficiency and model adaptability. Unlike the standard or inductive SVM, transductive SVM is formulated to find an optimal hyperplane to maximize the soft margin between positive and negative objects as well as between training and test data. It has also been theoretically proved that if properly tuned, transductive SVM generally performs no worse than its inductive counterpart (Wang et al. 2007).

For those reasons, we will use transductive SVM as the classifier. In the "Experiments" section, we will show how transductive SVM compares with inductive SVM.

Aspect-based Coherence Modeling

After aspects are recognized for each sentence, we then model text coherence from a topical perspective. Topics are organizational units that a human writer chooses and arranges to deliver a coherent train of thought. Modeling coherence thus hinges on modeling topic formation and transitions. We follow (Barzilay and Lee 2004) by using an HMM model with topics as states and sentences as observed sequences. But unlike their model that represents topics on the word level, we use aspects as semantic components of a topic, about which specific words are chosen. Figure 1 illustrates the difference between their model and ours with sentence generation mediated by aspects. The introduction of aspects contributes to a more intuitive modeling of the human writing process.

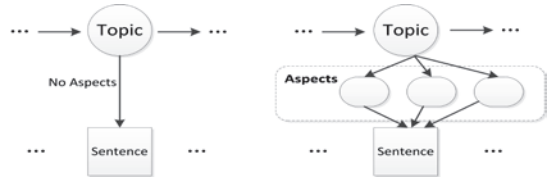

Figure 1: Models without (left) and with (right) aspects

Topic Induction

In our model, the topics are represented by sentence clusters. Like (Barzilay and Lee 2004), we use complete-link hierarchical clustering to cluster sentences. But unlike their work, we vectorize sentences using both words and aspects. The aspects are twice as much weighted as the words, which corresponds to the aspect's conceptual significance and leads to better clustering quality[2].

Among the clusters, one is merged from all smaller clusters (cluster size < M) that possibly contain non-essential information. We denote this merged cluster as c_0.

HMM Parameter Estimation

Given topics c_0, c_1, … with their corresponding HMM states s_0, s_1, …, we now estimate the HMM parameters. With no prior knowledge about the topics, we assume uniform distribution for the **state probabilities**. We denote aspects as a_i's and words as w_i's. Given a sentence $x = w_1 w_2 ... w_n$ having aspects: $\{a_1, ..., a_m\}$ and state s ($s \neq s_0$), the **emission probability** $P(x|s)$, shorthanded $P_s(x)$, is:

$$P_s(x) = \sum_{i=1}^{m} P_s(x \mid a_i) P_s(a_i) \qquad (1)$$

For aspect $a \in A$, the set of all aspects, MLE is used to estimate the raw probability of $P_s{}^*(a) : (s \neq s_0)$

$$P_s{}^*(a) = (Count_c(a) + \delta_1) / (\sum_{a'} Count_c(a') + \delta_1 \mid A \mid)$$

where $Count_c(a)$ is the count of a in cluster c (corresponding to s) and δ_1 is a smoothing coefficient. Note that some sentences may not have aspects and in this case, we use a special a_0 to represent the "empty aspect" and: ($s \neq s_0$)

$$P_s{}^*(a_0) = \prod_{i=1}^{|A|} (1 - P_s{}^*(a_i))$$

The raw probabilities are normalized so that they sum up to 1: ($s \neq s_0$) $P_s(a) = P^*{}_s(a) / \sum_{a'} P^*{}_s(a')$

For $P_{s_0}(a)$, we make it complementary to the other $P_s(a)$'s, as in (Barzilay and Lee 2004):

$$P_{s_0}(a) = (1 - Max_{s' \neq s_0} P_{s'}(a)) / \sum_{a' \in A \cup \{a_0\}} (1 - Max_{s' \neq s_0} P_{s'}(a'))$$

$P_s(x|a)$ in (1) can be estimated by taking the aspect-conditioned word generation and a bigram language model:

$$P_s(x \mid a) = P_s(w_{1...} w_n \mid a) \approx \prod_{i=1}^{n} (P_s(w_i \mid a) + P_s(w_i \mid w_{i-1}))$$

[2] We evaluated different sentence vectorizing schemes using the Silhou ette (Rousseeuw 1987) and Rand (Rand 1971) measures.

$$P_s(w \mid a) = (Count_c(w(a)) + \delta_2) / (Count_c(a) + \delta_2 \mid V \mid)$$

where $Count_c(w(a)) = \mid \{a' : w \in s \wedge s \supset a' \wedge a' \in c\} \mid$, $s \neq s_0$, and V is the vocabulary. For $P_{s_0}(w \mid a)$,

$$P_{s_0}(w \mid a) = (1 - Max_{s' \neq s_0} P_{s'}(w \mid a)) / \sum_{w' \in V} (1 - Max_{s' \neq s_0} P_{s'}(w' \mid a'))$$

We use the Bayesian rule for a_0:

$$P_s{*}(w \mid a_0) = P_s(w) P(a_0 \mid w) / P_s(a_0)$$

$$= P_s(w) \prod_{i=1}^{p} (1 - \frac{P_s(w \mid a_i) P(a_i)}{P_s(w)}) / P_s(a_0) \qquad (2)$$

and after normalization,

$$P_s(w \mid a_0) = P{*}_s(w \mid a_0) / \sum_{w'} P{*}_s(w' \mid a_0)$$

To calculate $P_s(w)$ in (2), for $s \neq s_0$,

$$P_s(w) = (Count_c(w) + \delta_3) / (\sum_{w' \in V} Count_c(w') + \delta_3 \mid V \mid)$$

$$P_{s_0}(w) = (1 - Max_{s' \neq s_0} P_{s'}(w)) / \sum_{w' \in V} (1 - Max_{s' \neq s_0} P_{s'}(w'))$$

The estimation of $P_s(w|w')$ is as in (Barzilay and Lee 2004) and then we have:

$$P_s{*}(x \mid a) = \prod_{i=1}^{n} (P_s(w_i \mid a) + P_s(w_i \mid w_{i-1}))$$

After normalization,

$$P_s(x \mid a) = P{*}_s(x \mid a) / \sum_{x'} P{*}_s(x' \mid a)$$

The state **transition probabilities** are estimated from two sources: sentences ($P_{sent}(s_j|s_i)$) and aspects ($P_{aspect}(s_j|s_i)$).

$$P_{sent}(s_j \mid s_i) = (SC(c_i, c_j) + \delta_4) / (SC(c_i) + \delta_4 r)$$

$$P_{aspect}(s_j \mid s_i) = (AC(c_i, c_j) + \delta_5) / (\sum_{j=1}^{r} AC(c_i, c_j) + \delta_5 r)$$

where r is the total number of topics (states), $SC(c, c')$ represents the count of documents where a sentence from c immediately precedes a sentence from c', $SC(c)$ represents the total count of documents with sentences from c. $AC(c, c')$ represents the count of documents where a sentence from c contains an aspect that immediately precedes an aspect contained in a sentence from c'. Aspect precedence is estimated by aspect-bearing sentence precedence.

We estimate the sentence-based state transitions and the aspect-based state transitions differently because unlike sentences, aspects are not unique in a document. The final transition probability is a linear combination of them:

$$P(s_j \mid s_i) = \lambda_1 P_{sent}(s_j \mid s_i) + (1 - \lambda_1) P_{aspect}(s_j \mid s_i)$$

where λ_1 is a coefficient in 0 .. 1.

Parameter Re-estimation and Coherent Ordering Determination

The original sentence clustering does not account for sentence order information in the training data. To utilize this important information for sentence ordering, we re-cluster the sentences by assigning each one to the topic (state) that most likely emits it, determined by Viterbi decoding. Then the HMM parameters are re-estimated and we iterate the process until clusters converge (Barzilay and Lee 2004). With a learned HMM model, we can determine the most coherent sentence ordering by selecting among all possible permutations one with the highest likelihood, computed by the forward algorithm.

Aspect-guided Summarization

To do extractive summarization, we build an aspect-guided summarizer following the pipeline of sentence selection and sentence ordering. Aspect information plays a significant role in both steps.

Sentence Selection with a Base Summarizer

We first describe an aspect-agnostic summarizer using a simple method (Zhang et al. 2011). The following formula is used to calculate the frequency score of a sentence s in document set D.

$$freq\_score(s) = \frac{\sum_{w \in s} TF_s(w) \cdot score(w)}{\sum_{w \in s} TF_s(w) \cdot ISF(w)} \qquad (3)$$

where $score(w) = \log TF_D(w)$, and the word w is a frequent or document topic word (i.e., a word used in the description of a document set); otherwise $score(w) = 0$. $ISF(w)$ is the inverted sentence frequency of w in the document set, defined as $ISF(w) = \log(N_s / SF_D(w))$. $TF_s(w)$ and $TF_D(w)$ are the frequencies of w in s and D; $SF_D(w)$ is the sentence frequency of w in D and N_s is the total number of sentences in D. The ISF-based sentence length is used to discount important words less.

Summary sentences are selected iteratively until summary length is reached. In each iteration, we select the top ranking sentence s* and then discount the frequency of all the words in s* by multiplying $\alpha < 1$. Redundant sentences are dropped using cosine similarity.

Sentence Selection with Recognized Aspects

Next we integrate the recognized sentential aspect information into the base summarizer.

For a sentence s, we first calculate its aspect score:

$$aspect\_score(s) = \sum_{asp \in s} classify\_score(asp),$$

where $classify\_score(asp)$ indicates the classification confidence for aspect asp. For our current scheme, it is the value calculated by the decision function trained from transductive SVM.

The final score of a sentence is a linear combination of its frequency score and aspect score.

$$score(s) = \lambda_2 \times freq\_score(s) + (1 - \lambda_2) \times aspect\_score(s)$$

where λ_2 is a coefficient in 0 .. 1. The iterative sentence selection algorithm is similar to that described for the base summarizer. The main difference is that after each iteration, not only the word scores but also the aspect scores are updated. For all the aspects in a selected sentence s*, $classify\_score(asp) = \beta \cdot classify\_score(asp)$, $\beta < 1$.

Sentence Ordering for Aspect-based Coherence

After we select all the sentences that meet the summary length requirement, we order them by considering all possible sentence permutations. Since aspect-guided summaries and source documents obviously differ in aspect density and content structure, we train an HMM model with aspect-annotated human summaries for similar documents. Then we select the best ordering among all sentence permutations as the sequence with the highest likelihood according to the HMM model parameters. This straightforward approach integrates well into the selection-ordering scheme. In the next section, we will show the efficacy of our simple method, especially for coherence enhancement.

We should also point out that for multi-document summarization, the summarization strategy in (Barzilay and Lee 2004), which attempts to correlate summary sentences with source sentences, cannot be adopted because it only works for single-document summarization. It is also pointless to train an HMM model with sentences simultaneously from different documents.

Experiments

We evaluate our method on three tasks: aspect recognition, text ordering, and summarization.

Evaluating Aspect Recognition

Our experimental data are composed of TAC 2010 source documents. For each of the documents of a category (*accidents*, *attacks*, *health and safety*, *resources*, *investigations*), we annotated a predefined list of aspects[3] for each sentence. Each of the 5 categories contains approximately 2000 sentences, from with 90% are used for training and the rest for test. Table 2 lists the aspects of category D3 (*health and safety*) and its aspects with brief explanations, followed by a sample annotated sentence.

D3.1 *WHAT*	what is the issue
D3.2 *WHO AFFECTED*	who is affected
D3.3 *HOW*	how they are affected
D3.4 *WHY*	why the issue occurs
D3.5 *COUNTERMEASURES*	countermeasures

Table 2: Aspects for Category D3 (*health and safety*)

The drugs were withdrawn in September 1997 after a Mayo Clinic study linked fen-phen to potentially fatal heart valve damage. {**D3.1**, **D3.3**, **D3.5**}

Implementation Details

To extract meta-phrase features, we use the Stanford Parser (Klein and Manning 2003) to do dependency parsing and extract all syntactico-semantic features. We use the OpenNLP tools[4] to find named entities for name-neighbor features. Features that occur only once are filtered. The inductive and transductive SVM classifications are implemented by using the SVM^{light} tool[5] with a linear kernel.

Results

For limit of space, we show the result only on one category (D3) in Figure 2, using different features. The results on the other categories are similar. The classifier is inductive SVM and we report the ten-fold cross validated F-measures on each aspect.

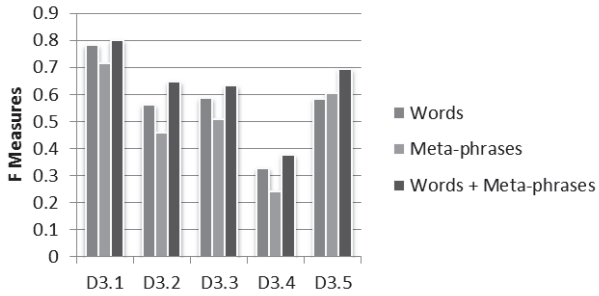

Figure 2: Using different features on *health and safety*

In most cases, the meta-phrase features help to improve classification performance, especially on aspects that may not be literally expressed (e.g., D3.3, D3.4, D3.5).

To test the multi-label classification and semi-supervised scheme, for each category we randomly select a small training set of 100 sentences as labeled data and 1500 different sentences as unlabeled data. Both word and meta-phrase features are used.

We compare both multi-class transformations (BD vs. LC) and classification algorithms (inductive SVM vs. transductive SVM). The evaluation metric is macro-average F measure, i.e., the average of F-measures on individual aspects. Figure 3 shows the aggregate result.

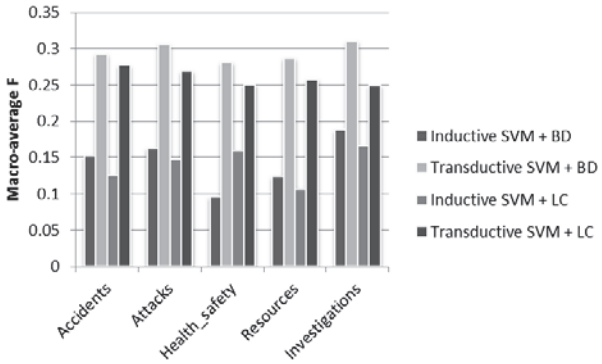

Figure 3: Macro-average F on the five categories

Transductive SVM defeats inductive SVM with an obvious advantage, showing the feasibility of semi-supervised

[3] See http://www.nist.gov/tac/2010/Summarization for details about the 5 categories and a total of 30 aspects.

[4] http://opennlp.sourceforge.net/

[5] http://svmlight.joachims.org/

learning for our task. In most cases, binary decomposition is superior to label combination.

Evaluating Text Ordering

Barzilay and Lee (2004) have shown the superiority of the HMM model in ordering to a baseline bigram model and a different probabilistic ordering model (Lapata 2003). Due to limit of space, we report how our aspect-based HMM model compares with their aspect-agnostic model.

Implementation Details

To evaluate the HMM model on summary text ordering, we annotated all the TAC 2010 human summaries for training (80%) and development (20%) data, with an average of 36.8 documents per category. We also annotated all the TAC 2011 human summaries for test, with an average of 35.2 documents per category[6]. We tune the HMM model parameters (M, δ_1 to δ_5, and λ_1) on the development data, as well as the number of topics (states).

Results

Figure 4 shows the difference between using aspect information and only literal information (NoAspect), measured by Kendall's τ, a widely used sequence ordering measure (Lapata 2006).

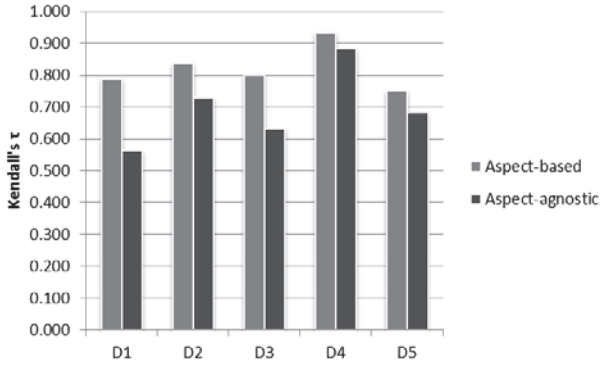

Figure 4: Ordering with and without aspects

Augmented with aspect information, the HMM model is shown to better represent the pattern of coherent ordering.

Evaluating Summarization

We evaluate the proposed summarization method on the TAC 2011 datasets (initial summarization subtask). According to our approach, aspect information is used in two stages: selecting sentences to cover particular information and ordering the selected sentences to enhance coherence. We use the classification model trained from all the TAC 2010 source documents to recognize aspects and the HMM model trained from all the TAC 2010 human summaries to order sentences selected for summary. All the model details and parameters are derived from the previous experiments on aspect recognition and text ordering.

Results

Information coverage is evaluated by the standard ROUGE measures (Lin and Hovy 2003). In Table 3, "Top" is the top ranking participant in TAC 2011 and "Average" is the average over all 50 TAC participants. Note that it is unnecessary to compare sentence orderings here because ROUGE is ordering-insensitive.

	ROUGE 2	*ROUGE SU4*
Base Summarizer	0.1206	0.1570
Base Summarizer + Aspect	**0.1223**	**0.1581**
Top	0.1337	0.1636
Average	0.0932	0.1266

Table 3: ROUGE evaluations of summaries

The base summarizer is a very competitive system (TAC ID: 4) in TAC 2011, ranking 5th and 4th in terms of ROUGE-2 and ROUGE-SU4, but it is outperformed by its aspect-enhanced version (TAC ID: 24) ranking 4th and 3rd in terms of ROUGE-2 and ROUGE-SU4. We observe that using recognized aspects helps to include more desirable information in summaries. But the improvement is limited, partly because the base summarizer has already included many aspects that happen to contain many frequent words.

To test the use of aspect in enhancing coherence, we employ two human judges to rate the coherence of summaries on a scale of 5 points, the higher the more coherent. For each TAC document set, we ask them to rate 4 summaries: 3 automatic summaries produced by the aspect-enhanced summarizer and 1 human summary. The automatic summaries differ from each other only in sentence ordering: following the selection sequence determined by sentence ranking scores (Ranking ordering), using the HMM model without aspect, i.e., Barzilay and Lee's (2004) model (BL ordering), using the HMM model with aspect (Aspect ordering). Cohen's Kappa is computed to be 0.71, indicating high inter-judge agreement. Table 4 lists the result, with the scores averaged over two judges.

Ranking ordering	*BL ordering*	*Aspect ordering*	*Human*
2.75	3.45	3.73	4.70

Table 4: Human Evaluation for Coherence

The differences between the two HMM ordering versions and the "Ranking ordering" or "Human" are very significant ($p < 0.0001$ on a paired two-tailed t-test). The difference between BL ordering and Aspect ordering is also significant ($p = 0.017$), though to a lesser degree. The 3.73 point by "Aspect ordering" proves that aspect-based ordering helps to generate fairly coherent summaries.

[6] We use only the initial summaries. There are 46 document sets for TAC 2010 and 44 for TAC 2011, each with 4 human summaries.

Conclusion and Future Work

Inspired by the TAC task, we propose a full-fledged approach to aspect-guided summarization. We recognize aspects in sentences with the help of the novel meta-phrase features and adapt an HMM model to aspect-based coherence. Based on sentence-level aspect information and the trained coherence model, we propose a simple but successful summarization model that leverage aspects in both sentence selection and ordering.

Our current model treats sentence selection and ordering as two independent modules. In the future, we will explore integrating the two modules to fully utilize annotated aspect information.

Acknowledgements

The work described in this paper was supported by the grant GRF PolyU 5230/08E.

References

Barzilay, R. and Lapata, M. 2008. Modeling Local Coherence: An Entity Based Approach. *Computational Linguistics*, 34:1 34.

Barzilay, R., and Lee, L. 2004. Catching the Drift: Probabilistic Content Models, with Applications to Generation and Summari zation. In *HLT NAACL 2004: Proceedings of the Main Confer ence*. 113 120.

Boutell, M. R., Luo, J., Shen, X., and Brown, C. M. 2004, Learn ing Multi label Scene Classification, *Pattern Recognition*, 37(9):1757 71.

Daumé III, H. and Marcu, D. 2006. Bayesian Query Focused Summarization. In *Proceedings of ACL 2006*, 305 312, Sydney, Australia.

Elsner, M., Austerweil, J. & Charniak E. 2007. "A Unified Local and Global Model for Discourse Coherence". In *Proceedings of NAACL HLT 2007*, 436 443. Rochester, NY.

Fuentes M, Alfonseca E, and Rodríguez H. 2007. Support Vector Machines for Query focused Summarization Trained and Evalu ated on Pyramid Data. In *Proceedings of the 45th annual meeting of the association for computational linguistics companion vol ume proceedings of the demo and poster sessions*, 57 60.

Ji, H., Favre, B., Lin, W., Gillick, D., Hakkani Tur, D., and Grishman, R. 2011. Open domain Multi document Summariza tion via Information Extraction: Challenges and Prospects. in *Multi source, Multilingual Information Extraction and Summari sation Volume of "Theory and Applications of Natural Language Processing"*. Springer.

Joachims, T. 1999. Transductive Inference for Text Classification using Support Vector Machines. In *Proceedings of the 16th Inter national Conference on Machine Learning (ICML)*.

Klein, D., and Manning, C. D. 2003. Accurate Unlexicalized Parsing. In *Proceedings of the 41st Meeting of the Association for Computational Linguistics*, 423 430.

Lapata, M. 2003. Probabilistic Text Structuring: Experiments with Sentence Ordering. In *Proceedings of the Annual Meeting of ACL*, 545 552. Sapporo, Japan.

Lapata, M. 2006. Automatic evaluation of information ordering: Kendall's tau. *Computational Linguistics*, 32(4):1 14.

Li, P., Wang, Y., Gao, W., and Jiang, J. 2011. Generating Aspect oriented Multi Document Summarization with Event aspect Model. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, 1137 1146. Edinburgh, Scotland, UK.

Li, W, Li, W, and Lu, Q. 2006. "Mining Implicit Entities in Que ries". In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC'06)*, 24 26. Genoa, Italy.

Lin, C Y. and Hovy, E. 2003. Automatic Evaluation of Summar ies Using N gram Co Occurrence Statistics. In *Proceedings of the Human Technology Conference 2003 (HLT NAACL 2003)*, 71 78, Edmonton, Canada.

McKnight, L., and Srinivasan, P. 2003. Categorization of Sen tence Types in Medical Abstracts. In *Proceedings of the Ameri can Medical Informatics Association Annual Symposium*, 440 444, Washington D.C.

Rand, W. M. 1971. Objective Criteria for the Evaluation of Clus tering Methods. *Journal of the American Statistical Association (American Statistical Association)*, 66 (336): 846 850.

Rousseeuw, P. J. 1987. Silhouettes: a Graphical Aid to the Inter pretation and Validation of Cluster Analysis. *Computational and Applied Mathematics* 20: 53 65.

Schilder, F. and Kondadadi, R. 2008. FastSum: Fast and Accurate Query based Multi document Summarization. In *Proceedings of ACL 08: HLT, short papers*, 205 208.

Teufel, S. and Moens, M. 1999. Argumentative Classification of Extracted Sentences as a First Step towards Flexible Abstracting. In I. Mani and M. T. Maybury (eds.), *Advances in Automatic Text Summarization*. 155 171. Cambridge, Massachusetts: MIT Press.

Teufel, S., and Moens, M. 2002. "Summarizing Scientific Arti cles: Experiments with Relevance and Rhetorical Status". *Com putational Linguistics*, 28(4): 409 445.

Tsoumakas, G. and Katakis, I. 2007. Multi label classification: An overview. *International Journal of Data Warehousing and Mining*, 3(3):1 13.

Vanderwende, L., Suzuki, H., Brockett, C., and Nenkova, A. 2007. "Beyond SumBasic: Task Focused Summarization with Sentence Simplification and Lexical Expansion". *Information Processing & Management* 43(6):1606 1618.

Vapnik, V. 1998. *Statistical Learning Theory*. New York: John Wiley & Sons.

Wan, X., Yang, J., and Xiao, J. 2007. Towards a Unified Ap proach Based on Affinity Graph to Various Multi document Summarizations. In *Proceedings of the 11th European conference*, 297 308.

Wang, L., Shen, X., and Pan, W. 2007. On Transductive Support Vector Machines. In J. Verducci, X. Shen, and J. Lafferty (eds.), *Prediction and Discovery*. American Mathematical Society.

Zhang, R., Ouyang, Y., and Li, W. 2011. Guided Summarization with Aspect Recognition. In *Proceedings of Textual Analysis Conference* (*TAC 2011*).

Zhou, L., Ticrea, M., and Hovy, E. 2004. Multidocument Biog raphy Summarization. In *Proceedings of Empirical Methods in Natural Language Processing* (*EMNLP 04*), 434 441.