Guided Summarization with Aspect Recognition

Renxian Zhang, You Ouyang, Wenjie Li
Department of Computing
The Hong Kong Polytechnic University
{csrzhang, csyouyang, cswjli}@comp.polyu.edu.hk}

1 Introduction

As a continuation of the summarization track of TAC 2010, the TAC 2011 summarization track pursues aspect-guided summarization. It is intended "to encourage a deeper linguistic (semantic) analysis of the source documents instead of relying only on document word frequencies to select important concepts". The sustained interest in guided summarization marks a significant turn to semantically oriented end results that implicitly favor deep (semantically rich) NLP, domain-specific IE, NLG, among other techniques. The update summarization task is similar to that included in earlier TAC summarization tracks.

The PolyCom (formerly PolyU) team has participated in the guided summarization task of both TAC 2010 and TAC 2011. Lacking experience with this new task, our team did not perform well with an IE-based system in 2010. This year, we draw on our learned lesson and resources built over months to build a system based on aspect recognition and a robust baseline. We achieve very competitive evaluation results released by NIST. In the following, we report the system design for our new system, including aspect recognition and aspect-guided summarization.

2 Aspect-bearing Sentence Recognition

Following the principle of guided summarization, we are committed to leveraging aspect information in our summarization system. In TAC 2010, we built a system that integrates aspect-bearing sentence recognition, sentential aspect recognition, and aspect-based sentence ranking. The sentence aspect recognition that relied on regular expression pattern induction, however, turned out to be computationally intractable and led to inferior performance.

For TAC 2011, we realize that as long as we can detect aspect-bearing sentences, sentential aspect recognition is not necessary for estimating the extract-worthiness of a sentence with aspect information. Besides, a system built on top of aspect-bearing sentence recognition and a robust frequency-based scheme is efficient and scalable. In the following, we provide more details of aspect-bearing sentence recognition as a text classification task.

2.1 Features

Traditionally, text classification tasks use word unigram features and represent objects as bag-of-words vectors. We believe, however, that words alone cannot deal competently with the linguistic richness of aspects, and 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 2 logical constituents and 11 grammatical roles. Their names and abbreviations are listed below.

Logical constituents			
predicate	PRED	argument	ARG
Grammatical roles			
nominal subject	nsubj	noun modifier	nn
controlling subject	xsubj	prepositional modifier	prepm
passive nominal subject	nsubjp	adjectival modifier	amod
direct object	dobj	agent	agent
indirect object	iobj	appositional modifier	appos
abbreviation modifier			abbrev

Table 1: Syntactic tags and their abbreviations

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 2 has examples.

NE types represent the semantic attributes of special NPs in a sentence, which are indicative of particular types of news details. We use 6 NE types: person (PER), organization (ORG), location (LOC), date (DAT), money (MON), and percentage (PCT). 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 2.

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 2: Meta-phrase patterns and examples

For each of the above textual pieces from our running example, 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 as shown in the following.

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 concept relations and NE contexts at different levels of abstraction. Different from previous dependency relation-based works (Nastase et al., 2006; Özgür and Güngör, 2010), we use not only dependency-related words, but also dependency relations per se and two higher-level constructs: PRED and ARG.

Name-neighbor meta-phrase extraction is a simple extension of NE recognition; syntactico-semantic meta-phrases are extracted in three scans as predicate-argument or argument-modifier relations are extracted 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.

2.2 Multi-label Classification

Since each sentence may be associated with an indefinite number of aspects, aspect recognition on the sentence level is a multi-label classification problem. According to the survey of Tsoumakas and Katakis (2007), problem transformation is a widely used strategy to tackle multi-label classification, which transforms multi-label classification to single-label classifications.

Two popular problem transformation methods are the label combination and binary decomposition methods (Boutell et al., 2004; Tsoumakas and Katakis, 2007). 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 sufficient training data may not be available for each transformed single-label class. Whereas binary decomposition (BD) assumes label independence which does not necessarily hold. In Section 2.5, we will show experimental results using both methods.

2.3 Transductive SVM

Up till the time of writing this report, we are not aware of any publically available large-sized training corpus for the specified aspects. Therefore we have to entertain two critical issues before performing classification: 1) insufficient training data may harm classification accuracy; 2) a model learned from limited training data may not adapt

well to unseen data. The second issue is raised because although NIST has assigned a fixed set of aspects to a specific category, that category may still contain highly diversified documents (Owczarzak, personal contact). For example, "health and safety" articles can range from Chinese food problems to the safety of a traffic device.

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.

For a classification problem $\{\mathbf{x}_i, y_i\}$ with $y_i \in \{+1, -1\}$, inductive SVM is formulated to find an optimal hyperplane $sign(\mathbf{w}\cdot\mathbf{x}_i - b)$ to maximize the soft margin between positive and negative objects, transductive SVM further considers test data $\mathbf{x}_i^*$ during training by finding a labeling y_j^* and a hyperplane to maximize the soft margin between both training and test data:

$$\text{minimize: } 1/2\|\mathbf{w}\|^2 + C_1\sum_i \phi_i + C_2\sum_i \varphi_i$$

$$\text{s.t. } y_i(\mathbf{x}_i \cdot \mathbf{w} - b) \geq 1 - \phi_i,\ \phi_i \geq 0$$

$$y_i^*(\mathbf{x}_i^* \cdot \mathbf{w} - b) \geq 1 - \varphi_i,\ \varphi_i \geq 0$$

where φ_i is a slack variable for the test data. In fact, labeling test data is done during training.

For those reasons, we will use transductive SVM to classify the TAC 2011 documents. In Section 2.5, we will show how transductive SVM compares with inductive SVM that we used in TAC 2010 in a particular experimental setting.

2.4 Training data

Starting from TAC 2010, we have manually constructed and maintained a training corpus for the guided summarization task. We selected news articles belonging to one of the five target categories ("accidents and natural disasters", "attacks", "health and safety", "endangered resources", and "investigations and trials") from past DUC/TAC test data. To the training data used in TAC 2010, we appended a new collection of documents from the officially released data of TAC 2010 with aspect information. For each of the articles of a certain category, we annotated the target aspects in the format of NIST-provided samples. Each of the five categories contains approximately 2000 sentences.

2.5 Evaluation of Classification

We did a pilot study of multi-class classification in order to decide an optimal classification scheme for summarization-oriented aspect recognition. We experimented on 2 categories: "health and safety" (H&S) and "trials and investigations" (T&I), with 5 and 6 aspects respectively. To simulate the real-life difficulties explained in Section 2.3, for each category we randomly select from our training corpus a small training set of 100 sentences as labeled data and a much larger test set of 1500 different sentences as unlabeled data.

We compared both multi-class transformations (BD vs. LC) and classification algorithms (inductive SVM vs. transductive SVM). We used the SVM^{light} tool [1] with a linear kernel. The evaluation metric is macro-average F measure. Table 3 shows the result.

Method	*H&S*	*T&I*
Inductive SVM + BD	0.096	0.152
Transductive SVM + BD	0.281	0.293
Inductive SVM + LC	0.159	0.125
Transductive SVM + LC	0.251	0.277

Table 3: Macro-average F on the two datasets

Obviously, transductive SVM demonstrates a pronounced advantage over inductive SVM with our experimental setting. The choice of classification algorithm is also the deciding factor of classification performance. In most cases, binary decomposition is also superior to label combination. According to such results, we will apply transductive SVM and binary decomposition on our training data to the unseen data released for the TAC 2011 summarization track.

3 Aspect-guided Summarization

It is noteworthy that in TAC 2010, many systems chose to bypass the complexities introduced by aspects by reformulating aspects as queries, fitting a topic model to the pre-defined aspects, or ignoring aspects altogether under the assumption that the extracted sentences will automatically contain the required aspects. Surprisingly, such aspect-agnostic systems were successful to certain extents (according to the TAC 2010 reports). By

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

contrast, our purely IE-driven aspect-oriented system performed poorly.

With such lessons learned from the previous year, this time we experiment with two systems: a robust but aspect-agnostic frequency-based baseline system and its aspect-integrated version.

3.1 Baseline System

We implement a frequency-based extraction model. The following formula is used to calculate the frequency score of a sentence *s*.

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

In this formula, $score(w)$ is an estimation of the word importance, calculated as $score(w) = \log TF_D(w)$, when w satisfies (1) $TF_D(w) > c$, (2) w is not a stop-word, (3) w is a category description word; otherwise $score(w) = 0$[2]. $ISF(w)$ is the inverted sentence frequency of w in the input document set, calculated as $ISF(w) = \log \frac{N_s}{SF_D(w)}$. $TF_s(w)$, $TF_D(w)$ are the frequencies of w in the sentence s and the input document set D respectively; $SF_D(w)$ is the sentence frequency of w in D and N_s is the total number of sentences in D.

Using the above formula, the process of extracting sentences with dynamic word scoring is as follows.

While the summary length does not exceed the word limit

Calculate the word importance by $score(w) = \log TF_D(w)$;

Rank the sentences by $\frac{\sum_{w_i \in s} TF_s(w) \cdot score(w)}{\sum_{w_i \in s} TF_s(w) \cdot ISF(w)}$;

Select the highest ranked sentence s_0;

Update the frequency of all the words appearing in s_0 by $TF_D(w) = \alpha \cdot TF_D(w)$; [3]

Figure 1: Algorithm of the baseline system

The ISF-based sentence length is used to give words different weights when counting sentence length. If a word is more dominant in the input document set, it should be considered shorter so that the sentence containing it should be penalized less by length.

To control sentence redundancy, we inspect every newly extracted sentence by comparing it with previously extracted sentences. If the new sentence is too similar to what are already selected, we drop it.

3.2 Aspect-integrated System

Next we integrate sentential aspect information learned from our training data, as described in Section 2, into the baseline system.

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

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

where *classify_score*(*asp*) indicates the classification confidence for aspect *asp*. For our current scheme, it is the classification score calculated by transductive SVM.

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

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

The summarization algorithm is similar to that presented in Figure 1. The main difference is that after each round of sentence selection, not only the word scores but also the aspect scores are updated.

Estimated on the TAC 2010 dataset, λ is decided to be 0.8, indicating the dominance of the frequency-based scheme and lending further credence to the success of aspect-agnostic methods as was pointed out in the beginning of this section.

3.3 Update Summarization

For this task, sentence novelty takes priority over aspect information. Therefore, we continue to adopt the simple method used in TAC 2010. First, the aspect-bearing sentences in document set B are recognized. Then they are ranked and selected according to word frequency and aspect coverage. In addition, we discard any sentence that is highly similar to any sentence in document set A of the

[2] c is a threshold which is empirically decided to be 5 .

[3] The damping factor α is empirically decided to be $1/e$.

same topic. In our implementation, sentence similarity is the cosine similarity between their term vectors and "highly similar" is translated to a value above a high threshold (0.75).

4 TAC 2011 Evaluation Results

In TAC 2011, the PolyCom team submitted two runs: PolyCom1 and PolyCom2. PolyCom1 is the baseline (aspect-agnostic) system and PolyCom2 is the aspect-integrated system. Their run IDs in the official evaluation report are #4 and #24. In the following, we present the official evaluation results by ROUGE, BE, and manual evaluation metrics.

4.1 ROUGE

Table 4 and 5 show the ROUGE-2 (R-2) and ROUGE-SU4 (R-SU4) scores of our systems, for both the original summaries and update summaries. For comparison, we list the top system and all system average (including the 2 baselines). We also show the ranks of the top system and our two runs among a total of 50 runs.

ID	*R-2/Rank*	*R-SU4/Rank*
#43	0.13440 / 1	0.16519 / 1
PolyCom2	0.12306 / 4	0.15975 / 3
PolyCom1	0.12133 / 5	0.15865 / 4
Average	0.09005	0.12729

Table 4: Original summary ROUGE

ID	*R-2/Rank*	*R-SU4/Rank*
#43	0.09581 / 1	0.1308 / 1
PolyCom2	0.08643 / 4	0.12803 / 2
PolyCom1	0.08507 / 6	0.12787 / 4
Average	0.07002	0.10940

Table 5: Update summary ROUGE

It is obvious that our aspect-integrated system performs very competitively. A little unexpectedly, the aspect-agnostic system (PolyCom1) performs well and not far behind its aspect-integrated cousin. On the one hand, such result shows the robustness of a good frequency-based scheme. On the other hand, it indicates aspect information is not so useful as predicted, at least measured by ROUGE, for the task design.

4.2 BE

Table 6 shows the BE results of our two systems, for both the original summaries (Ori) and update summaries (Upd), in the same format as before.

ID	*Ori/Rank*	*Upd/Rank*
#43	0.08565 / 1	0.06473 / 1
PolyCom2	0.07938 / 4	0.05437 / 9
PolyCom1	0.07702 / 5	0.05448 / 8
Average	0.05700	0.04251

Table 6: BE results

The BE metric relies on fine-grained matching between useful syntactic elements. Again, we achieve very competitive results with both the aspect-integrated version and the aspect-agnostic version, especially for the original summaries.

4.3 Manual evaluation metrics

The manual evaluations are aimed to complement the automatic evaluation metrics, measuring not only informativeness, but also expressiveness. Among them, the Pyramid score is based on the presence of human-annotated SCUs that correspond to aspects. Therefore, it might be a more appropriate metric for the aspect-guided task.

Table 7 shows the result of our systems, in a similar format as before.

ID	*Ori/Rank*	*Upd/Rank*
#22/#9	(#22) 0.471 / 1	(#9) 0.346 / 1
PolyCom2	0.437 / 8	0.3 / 17
PolyCom1	0.447 / 4	0.332 / 8
Average	0.367	0.267

Table 7: Pyramid results

Our systems continue to perform well, especially for the original summaries. But unexpectedly, the aspect-agnostic system outscores the aspect-integrated system, demonstrating again the robustness of a good frequency-based scheme.

The other two manual evaluation metrics are linguistic quality and overall responsiveness. Tables 8 and 9 show the results.

ID	*Ori/Rank*	*Upd/Rank*
#32/#1	(#32) 3.75 / 1	(#1) 3.455 / 1
PolyCom2	2.932 / 23	2.795 / 25
PolyCom1	2.886 / 26	2.795 / 25
Average	2.761	2.738

Table 8: Linguistic quality results

ID	*Ori/Rank*	*Upd/Rank*
#25/#35	(#25) 3.159 / 1	(#35) 2.591 / 1
PolyCom2	2.818 / 28	2.364 / 17
PolyCom1	2.955 / 16	2.523 / 5
Average	2.685	2.231

Table 9: Overall responsiveness results

Admittedly, our systems do not perform as well on those metrics as on the previous metrics, although they invariably outperform the average. One reason is that we have not accommodated any substantial readability-enhancing modules, since we are mainly concerned with the use of aspect information. It is our next step, however, to produce better organized, more focused, and more readable summaries by manipulating aspect relations and other related information. We hope to make breakthroughs in this direction, which is to be witnessed by the system report for TAC 2012.

5 Conclusion and Suggestion

The PolyCom team has built two systems for TAC 2011, one robust baseline system that makes no use of aspect, and one that extends the baseline system with sentential aspect information.

Acquiring sentential aspect information is our emphasis. We formulate the problem of recognizing aspect on the sentence level as a classification problem and develop a model that utilizes rich textual features and transductive SVM. The trained model is then used to predict aspect-bearing sentences and the predicted information is used to build an aspect-integrated system that is biased to both frequent and aspect-related information.

As evaluated by NIST, our two systems perform very competitively in terms of aspect coverage. In most cases, the aspect-agnostic baseline system is not much worse than the aspect-integrated system and sometimes outscores the latter. This fact entails two ramifications: 1) the frequency-based method proves to be a robust prototype for various summarization tasks; 2) the design of the guided summarization task cannot effectively discriminate aspect-aware methods from aspect-agnostic methods.

To address the second ramification, we suggest a modification to the task design: instead of proposing a seemingly comprehensive list of aspects for a certain category of news articles, the task designer may consider making a "biased" list of aspects. For example, instead of asking for "when", "where", and "what happened" for an accident, we may only ask for "countermeasures", "casualties", and even "public responses". The underlying assumption is that a frequency-based method, though aspect-agnostic, will "accidentally" select aspect-rich content if the target aspects are the most typical news element for a certain category, which are well captured by frequency. By biasing the aspect list, those frequency-based methods may not have the luck of discovering the less typical (or less frequent) news elements and the aspect-oriented methods may demonstrate their true advantage.

References

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

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

Nastase, V., Shirabad, J. S., and Caropreso, M. F. 2006. Using Dependency Relations for Text Classification. In *Proceedings of the Nineteenth Canadian Conference on Artificial Intelligence*, Quebec, Canada.

Özgür, L. and Güngör, T. 2010. Text Classification with the Support of Pruned Dependency Patterns. *Pattern Recognition Letters*, 31 (12):1598–1607.

Tsoumakas, G. and Katakis, I. 2007. Multi Label Classification: An Overview. *International Journal of Data Warehousing and Mining*, 3(3).

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