Extractive Summarization Using Supervised and Semi-supervised Learning

Kam-Fai Wong[*], Mingli Wu[*†]
[*]Department of Systems Engineering and Engineering Management
The Chinese University of Hong Kong
New Territories, Hong Kong
{kfwong,mlwu}@se.cuhk.edu.hk

Wenjie Li[†]
[†]Department of Computing
The Hong Kong Polytechnic University
Kowloon, Hong Kong
cswjli@comp.polyu.edu.hk

Abstract

It is difficult to identify sentence importance from a single point of view. In this paper, we propose a learning-based approach to combine various sentence features. They are categorized as surface, content, relevance and event features. Surface features are related to extrinsic aspects of a sentence. Content features measure a sentence based on content-conveying words. Event features represent sentences by events they contained. Relevance features evaluate a sentence from its relatedness with other sentences. Experiments show that the combined features improved summarization performance significantly. Although the evaluation results are encouraging, supervised learning approach requires much labeled data. Therefore we investigate co-training by combining labeled and unlabeled data. Experiments show that this semi-supervised learning approach achieves comparable performance to its supervised counterpart and saves about half of the labeling time cost.

1 Introduction

[1] Automatic text summarization involves condensing a document or a document set to produce a human comprehensible summary. Two kinds of summarization approaches were suggested in the past, i.e., extractive (Radev et al., 2004; Li et al., 2006) and abstractive summarization (Dejong, 1978). The abstractive approaches typically need to "understand" and then paraphrase the salient concepts across documents. Due to the limitations in natural language processing technology, abstractive approaches are restricted to specific domains. In contrast, extractive approaches commonly select sentences that contain the most significant concepts in the documents. These approaches tend to be more practical.

Recently various effective sentence features have been proposed for extractive summarization, such as signature word, event and sentence relevance. Although encouraging results have been reported, most of these features are investigated individually. We argue that it is ineffective to identify sentence importance from a single point of view. Each sentence feature has its unique contribution, and combing them would be advantageous. Therefore we investigate combined sentence features for extractive summarization. To determine weights of different features, we employ a supervised learning framework to identify how likely a sentence is important. Some researchers explored learning based summarization, but the new emerging features are not concerned, such as event features (Li et. al, 2006).

We investigate the effectiveness of different sentence features with supervised learning to decide which sentences are important for summarization. After feature vectors of sentences are examined, a supervised learning classifier is then employed. Particularly, considering the length of final summaries is fixed, candidate sentences are re-ranked. Finally, the top sentences are extracted to compile the final summaries. Experiments show that combined features improve summarization performance significantly.

Our supervised learning approach generates promising results based on combined features. However, it requires much labeled data. As this procedure is time consuming and costly, we investigate semi-supervised learning to combine labeled data and unlabeled data. A semi-

Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008), pages 985–992
Manchester, August 2008

supervised learning classifier is used instead of a supervised one in our extractive summarization framework. Two classifiers are co-trained iteratively to exploit unlabeled data. In each iteration step, the unlabeled training examples with top classifying confidence are included in the labeled training set, and the two classifiers are trained on the new training data. Experiments show that the performance of our semi-supervised learning approach is comparable to its supervised learning counterpart and it can reduce the labeling time cost by 50%.

The remainder of this paper is organized as follows. Section 2 gives related work and Section 3 describes our learning-based extractive summarization framework. Section 4 outlines the various sentence features and Section 5 describes supervised/semi-supervised learning approaches. Section 6 presents experiments and results. Finally, Section 7 concludes the paper.

2 Related Work

Traditionally, features for summarization were studied separately. Radev et al. (2004) reported that position and length are useful surface features. They observed that sentences located at the document head most likely contained important information. Recently, content features were also well studied, including centroid (Radev et al., 2004), signature terms (Lin and Hovy, 2000) and high frequency words (Nenkova e t al., 2006). Radev et al. (2004) defined centroid words as those whose average *tf***idf* score were higher than a threshold. Lin and Hovy (2000) identified signature terms that were strongly associated with documents based on statistics measures. Nenkova et al. (2006) later reported that high frequency words were crucial in reflecting the focus of the document.

Bag of words is somewhat loose and omits structural information. Document structure is another possible feature for summarization. Barzilay and Elhadad (1997) constructed lexical chains and extracted strong chains in summaries. Marcu (1997) parsed documents as rhetorical trees and identified important sentences based on the trees. However, only moderate results were reported. On the other hand, Dejong (1978) represented documents using predefined templates. The procedure to create and fill the templates was time consuming and it was hard to adapt the method to different domains.

Recently, semi-structure events (Filatovia and Hatzivassiloglou, 2004; Li et al., 2006; Wu, 2006) have been investigated by many researchers as they balanced document representation with words and structures. They defined events as verbs (or action nouns) plus the associated named entities. For instance, given the sentence "Yasser Arafat on Tuesday accused the United States of threatening to kill PLO officials", they first identified "accused", "threatening" and "kill" as event terms; and "Yasser Arafat", "United States", "PLO" and "Tuesday" as event elements. Encouraging results based on events were reported for news stories.

From another point of view, sentences in a document are somehow connected. Sentence relevance has been used as an alternative means to identify important sentences. Erkan and Radev (2004) and Yoshioka (2004) evaluate the relevance (similarity) between any two sentences first. Then a web analysis approach, PageRank, was used to select important sentences from a sentence map built on relevance. Promising results were reported. However, the combination of these features is not well studied. Wu et al. (2007) conducted preliminary research on this problem, but event features were not considered.

Normally labeling procedure in supervised learning is very time consuming. Blum and Mitchell (1998) proposed co-training approach to exploit labeled and unlabeled data. Promising results were reported from their experiments on web page classification. A number of successful studies emerged thereafter for other natural language processing tasks, such as text classification (Denis and Gilleron, 2003), noun phrase chunking (Pierce and Cardie, 2001), parsing (Sarkar, 2001) and reference or relation resolution (Muller et al., 2001; Li et al., 2004). To our knowledge, there is little research in the application of co-training techniques to extractive summarization.

3 The Framework for Extractive Summarization

Extractive summarization can be regarded as a classification problem. Given the features of a sentence, a machine-learning based classification model will judge how likely the sentence is important. The classification model can be supervised or semi-supervised learning. Supervised approaches normally perform better, but require more labeled training data. SVMs perform well in many classification problems. Thus we employ it for supervised learning. For semi-supervised learning, we co-trained a probabilistic

SVM and a Naïve Bayesian classifier to exploit unlabeled data.

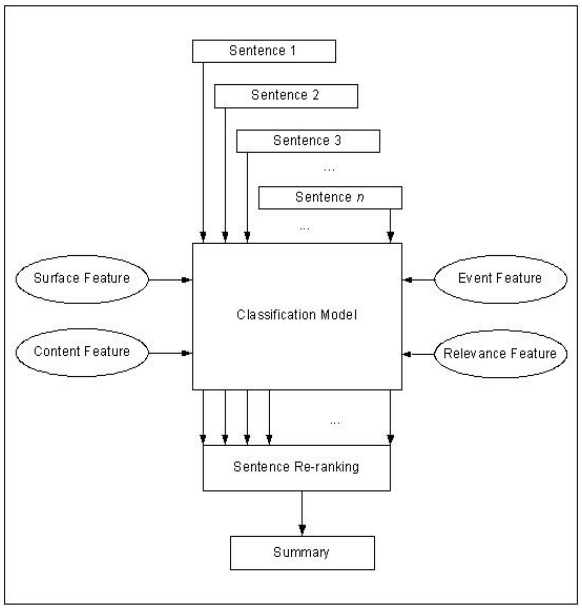

Figure 1. Learning-based Extractive Summarization Framework

The automatic summarization procedure is shown in Figure 1. First, each input sentence is examined by going through the pre-specified feature functions. The classification model will then predict the importance of each sentence according to its feature values. A re-ranking algorithm is then used to revise the order. Finally, the top sentences are included in the summaries until the length limitation is reached. The re-ranking algorithm is crucial, as more important content are expected to be contained in the final summary with fixed length. Important sentences above a threshold are regarded as candidates. The one with less words and located at the beginning part of a document is ranked first. The re-ranking algorithm is described as follows.

$Rank_i = RankPos_i + RankLength_i$

where $RankPos_i$ is the rank of sentence i according to its position in a document (i.e. the sentence no.) and $RankLength_i$ is rank of sentence i according to its length.

4 Sentence Features for Extractive Summarization

This section provides a detailed description on the four types of sentence features, i.e., surface, content, event and relevance features, which will be examined systematically.

4.1 Surface Features

Surface features are based on structure of documents or sentences, including sentence position in the document, the number of words in the sentence, and the number of quoted words in the sentence (see Table 1).

Name	Description
Position	1/sentence no.
Doc_First	Whether it is the first sentence of a document
Para_First	Whether it is the first sentence of a paragraph
Length	The number of words in a sentence
Quote	The number of quoted words in a sentence

Table 1. Types of surface features

The intuition with respect to the importance of a sentence stems from the following observations: (1) the first sentence in a document or a paragraph is important; (2) the sentences in the earlier parts of a document is more important than sentences in later parts; (3) a sentence is important if the number of words (except stop words) in it is within a certain range; (4) a sentence containing too many quoted words is unimportant.

4.2 Content Features

We integrate three well-known sentence features based on content-bearing words i.e., centroid words, signature terms, and high frequency words. Both unigram and bigram representations have been investigated. Table 2 summarizes the six content features we studied.

Name	Description
Centroid_Uni	The sum of the weights of centroid uni-gram
Centroid_Bi	The sum of the weights of centroid bi-grams
SigTerm_Uni	The number of signature uni-grams
SigTerm_Bi	The number of signature bi-grams
FreqWord_Uni	The sum of the weights of frequent uni-grams
FreqWord_Bi	The sum of the weights of frequent bi-grams

Table 2. Types of content features

4.3 Event Features

An event is comprised of an event term and associated event elements. In this study, we choose verbs (such as “elect and incorporate”) and action nouns (such as “election and incorporation”) as event terms that can characterize actions. They relate to “did what”. One or more associated named entities are considered as event elements. Four types of named entities are currently under

consideration. The GATE system (Cunningham et al., 2002) is used to tag named entities, which are categorized as <Person>, <Organization>, <Location> and <Date>. They convey the information about "who", "whom", "when" and "where". A verb or an action noun is deemed an event term only when it appears at least once between two named entities.

Event summarization approaches based on instances or concepts are investigated. An occurrence of an event term (or event element) in a document is considered as an instance, while the collection of the same event terms (or event elements) is considered as a concept. Given a document set, instances of event terms and event elements are identified first. An event map is then built based on event instances or concepts (Wu , 2006; Li et al., 2006). PageRank algorithm is used to assign weight to each node (an instance or concept) in the event map. The final weight of a sentence is the sum of weights of event instances contained in the sentence.

4.4 Relevance Features

Relevance features are incorporated to exploit inter-sentence relationships. It is assumed that: (1) sentences related to important sentences are important; (2) sentences related to many other sentences are important. The first sentence in a document or a paragraph is important, and other sentences in a document are compared with the leading ones. Two types of sentence relevance, FirstRel_Doc and FirstRel_Para (see Table 3), are measured by comparing pairs of sentences using word-based cosine similarity.

Another way to exploit sentence relevance is to build a sentence map. Every two sentences are regarded relevant if their similarity is above a threshold. Every two relevant sentences are connected with a unidirectional link. Based on this map, PageRank algorithm is applied to evaluate the importance of a sentence. These relevance features are shown in Table 3.

Name	Description
FirstRel_Doc	Similarity with the first sentence in the document
FirstRel_Para	Similarity with the first sentence in the paragraph
PageRankRel	PageRank value of the sentence based on the sentence map

Table 3. Types of relevance features

5 Supervised/Semi-supervised Learning Approaches

To incorporate features described in Section 4, we investigate supervised and semi-supervised learning approaches. Probabilistic Support Vector Machine (PSVM) is employed as supervised learning (Wu et al., 2004), while the co-training of PSVM and Naïve Bayesian Classifier (NBC) is used for semi-supervised learning. The two learning-based classification approaches, PSVM and NBC, are described in following sections.

5.1 Probabilistic Support Vector Machine (PSVM)

For a set of training examples (x_i , y_i), $i = 1,..., l$, where x_i is an instance and y_i the corresponding label, basic SVM requires the solution of the following optimization problem.

$$\min_{w,b,\xi} \ \frac{1}{2} w^T w + C \sum_{i=1}^{l} \xi_i$$

$$\text{subject to} \quad y_i (w^T \phi(x_i) + b) \geq 1 - \xi_{i,} \qquad \xi_i \geq 0$$

Here the SVM classifier is expected to find a hyper-plane to separate testing examples as positive and negative. Wu et al. (2004) extend the basic SVM to a probabilistic version. Its goal is to estimate

$$p_i = p(y = i \mid x), i = 1,...k .$$

First the pairwise (one-against-one) probabilities $r_{ij} \approx p(y = i \mid y = i \text{ or } j, x)$ is estimated using

$$r_{ij} \approx \frac{1}{1 + e^{Af + B}}$$

where *A* and *B* are estimated by minimizing the negative log-likelihood function using training data and their decision values *f*. Then p_i is obtained by solving the following optimization problem

$$\min_{p} \ \frac{1}{2} \sum_{i=1}^{k} \sum_{j: j \neq i} (r_{ji} p_i - r_{ij} p_j)^2$$

$$\text{subject to} \quad y_i (w^T \phi(x_i) + b) \geq 1 - \xi_{i,} \qquad \xi_i \geq 0$$

The problem can be reformulated as

$$\min_{P} \ \frac{1}{2} P^T Q P$$

where $Q_{ij} = \begin{cases} \sum_{s:s\neq i} r_{si}^2 & \text{if } i = j \\ -r_{ji}r_{ij} & \text{if } i \neq j \end{cases}$

The problem is convex and the optimality conditions a scalar *b* such that

$$\begin{bmatrix} Q & e \\ e^T & 0 \end{bmatrix}\begin{bmatrix} \mathrm{P} \\ \mathrm{b} \end{bmatrix} = \begin{bmatrix} \mathrm{z} \\ 1 \end{bmatrix}$$

where *e* is the vector of all 1s and *z* is the vector of all 0s, and *b* is the Lagrangian multiplier of the equality constraint $\sum_{i=1}^{k} p_i = 1$.

5.2 Naïve Bayesian Classier (NBC)

Naïve Bayesian Classier assumes features are independent. It learns prior probability and conditional probability of each feature, and predicts the class label by highest posterior probability. Given a feature vector (F_1, F_2, F_3,…, F_n), the classifier need to decide the label *c*:

$$c = \arg\max_c P(c \mid F_1, F_2, F_3, ...F_n)$$

By applying Bayesian rule, we have

$$P(c \mid F_1, F_2, F_3, ..., F_n) = \frac{P(c)P(F_1, F_2, F_3, ..., F_n \mid c)}{P(F_1, F_2, F_3, ..., F_n)}$$

Since the denominator does not depend on *c* and the values of F_i are given, therefore the denominator is a constant and we are only interested in the numerator. As features are assumed independent,

$$P(c \mid F_1, F_2, F_3, ...F_n) = P(c)P(F_1, F_2, F_3, ...F_n \mid c)$$
$$\approx P(c)\prod_{i=1}^{n} P(F_i \mid c)$$

where $P(F_i \mid c)$ is estimated with MLE from training data with Laplace Smoothing.

5.3 Co-Training (COT)

Supervised learning approaches require much labeled data and the labeling procedure is very time-consuming. Literature (Blum and Mitchell, 1998; Collins, 1999) has suggested that unlabeled data can be exploited together with labeled data by co-training two classifiers. (Blum and Mitchell, 1998) trained two classifiers of same type on different features, and (Li et al., 2004) trained two classifiers of different types. In this paper, as the number of involved features is not too many, we train two different classifiers, PSVM and NBC, on the same feature spaces. The co-training algorithm is described as follows.

Given:
L is the set of labeled training examples
U is the set of unlabeled training examples
Loop: until the unlabeled data is exhausted
Train the first classifier C_1 (PSVM) on L
Train the second classifier C_2 (NBC) on L
For each classifier C_i
C_i labels examples from U
C_i chooses *p* positive and *n* negative examples E from U. These examples have top classifying confidence.
C_i removes examples E from U
C_i adds examples E with the corresponding labels to L
End
Output: label the test examples by the optimal classifier which is evaluated on training data according to the classification performance.

6 Experiments

DUC 2001[2] has been used in our experiments. It contains 30 clusters of relevant documents and 308 documents in total. Each cluster deals with a specific topic (e.g. a hurricane) and comes with model summaries created by NIST assessors. 50, 100, 200 and 400 word summaries are provided. Twenty-five of the thirty document clusters are used as training data and the remaining five are used as testing. The training/testing configuration is same in experiments of supervised learning and semi-supervised learning, while the difference is that some sentences in training data are not tagged for semi-supervised learning.

An automatic evaluation package, i.e., ROUGE (Lin and Hovy, 2003) is employed to evaluate the summarization performance. It compares machine-generated summaries with model summaries based on the overlap. Precision and recall measures are used to evaluate the classification performance. For comparison, we evaluate our approaches on DUC 2004 data set also. It contains 50 clusters of documents. Only 665-character summaries are given by assessors for each cluster.

6.1 Experiments on Supervised Learning Approach

We use LibSVM[3] as our classification model for SVM classifiers normally perform better. Types of features presented in previous section are evaluated individually first. Precision measures

[2] http://duc.nist.gov/
[3] http://www.csie.ntu.edu.tw/~cjlin/libsvm/

the percentage of true important sentences among all important sentences labeled by the classifier. Recall measures the percentage of true important sentences labeled by the classifier among all true important sentences.

Table 4 shows the precisions and recalls of different feature groups under the PSVM classifier. Table 5 records the ROUGE evaluation results – ROUGE-1, ROUGE-2 and ROUGE-L. They evaluate the overlap between machine-generated summaries and model summaries based on unigram, bigram and long distance respectively. The summary length is limited to 200 words here.

Feature	Precision	Recall
Sur	0.488	0.146
Con	0.407	0.167
Rel	0.488	0.146
Event	0.344	0.146
Sur+Con	0.575	0.160
Sur+Rel	0.488	0.146
Con+Rel	0.588	0.139
Sur+Event	0.600	0.125
Con+Event	0.384	0.194
Rel+Event	0.543	0.132
Sur+Con+Event	0.595	0.153
Sur+Rel+Event	0.553	0.146
Con+Rel+Event	0.581	0.125
Sur+Con+Rel	**0.595**	**0.174**
Sur+Con+Rel+Event	0.579	0.153

Table 4. Classification performance based on different feature groups

Feature	Rouge-1	Rouge-2	Rouge-L
Sur	0.373	0.103	0.356
Con	0.352	0.074	0.334
Rel	0.373	0.103	0.356
Event	0.344	0.064	0.325
Sur+Con	0.380	0.109	0.363
Sur+Rel	0.373	0.103	0.356
Con+Rel	0.375	0.103	0.358
Sur+Event	0.348	0.091	0.332
Con+Event	0.344	0.071	0.330
Rel+Event	0.349	0.089	0.356
Sur+Con+Event	0.379	0.106	0.363
Sur+Rel+Event	0.371	0.101	0.353
Sur+Con+Rel	**0.396**	**0.116**	**0.358**
Sur+Con+Rel+Event	0.375	0.106	0.359

Table 5. ROUGE evaluation results for different feature groups

From Table 4, we can see the most useful feature groups are "surface" and "relevance", i.e. the external characteristics of a sentence in the document and the relationships of a sentence with other sentences in a cluster. The evaluation scores from surface features and relevance features are the same. We found that the reason is that the dominating feature in each feature group is about whether a sentence is the first sentence in a document. The influence of event features is not very positive. Based on our analysis the reason is that not all clusters contain enough event terms/elements to build a good event map.

From Table 5, it can be seen that the combination of multiple features or multiple feature groups outperforms individual feature or feature groups. When surface, content and relevance features are employed, the best performance is achieved, i.e., ROUGE-1 and ROUGE-2 score are 0.396 and 0.116 respectively. In our preliminary experiments, we find ROUGE-1 score of a model summary is 0.422 (without stemming and filtering stop words). Therefore summaries generated by our supervised learning approach received comparable performance with model summaries when evaluated by ROUGE. Although ROUGE is not perfect at this time, it is automatic and good complement to subjective evaluations.

We also find that the Rouge scores are similar for variations on the feature set. Sentences from original documents are selected to build the final summaries. Normally, only four to six sentences are contained in one 200-word summary in our experiments, i.e., few sentences will be kept in a summary. As variations of the feature set only induce little change of the order of most important sentences, the ROUGE scores change little.

6.2 Experiments on Semi-supervised Learning Approach

Supervised learning approaches normally achieve good performance but require manually labeled data. Recent literature (Blum and Mitchell, 1998; Collins, 199) has suggested that co-training techniques reduce the amount of labeled data. They trained two homogeneous classifiers based on different feature spaces. However this method is unsuitable for our application as the number of required features in our case is not too many. Therefore we develop a co-training approach to train different classifiers based on same feature space. PSVM and NBC are applied to the combination of surface, content and relevance features.

The capability of different learning approaches to identify important sentences is shown in Fig-

ure 2. The "x" axis shows the number of labeled sentences employed. The remained training sentences in DUC 2001 are employed as unlabeled training data. The y axis shows f-measures of important sentences identified from the test set. The size of the training seed set is investigated. For each size, three different seed sets which are chose randomly are used. The average evaluation scores are used as the final performance. This procedure avoids the variance of the final evaluation results. The ROUGE evaluation results of these supervised learning approaches and semi-supervised learning approaches are shown in Table 6 (2000 labeled sentences). It can be seen that the ROUGE performance of co-trained classifiers is better than that of individual classifiers.

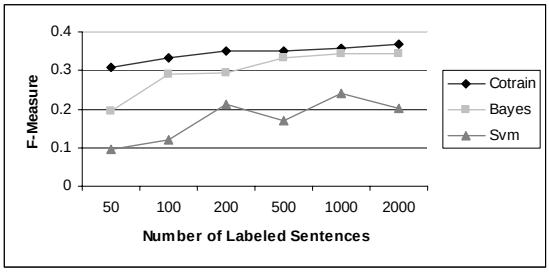

Figure 2. Performance of supervised learning and semi-supervised learning approaches

Learning Approaches	Rouge-1	Rouge-2	Rouge-L
PSVM	0.358	0.082	0.323
NBC	0.353	0.061	0.317
COT	**0.366**	**0.090**	**0.329**

Table 6. ROUGE evaluation results of supervised learning and semi-supervised learning

6.3 Experiments on Summary Length

In DUC 2001 dataset, 50, 100, 200 and 400-word summaries are provided to evaluate summaries with different length. Our supervised approach, which generates the best performance in previous experiments, is employed. The ROUGE scores of evaluations on different summary length are shown in Table 7. Our summaries consist of extracted sentences. It can be seen that these summaries achieve lower ROUGE scores when the length of summary is reduced. The reason is that when people try to write a more concise summary, condensed contents are included in the summaries, which may not use the original contents directly. Therefore the word-overlapping test tool in ROUGE generates lower scores.

We then tested the same classifier and same features on DUC 2004. The length of summaries is only 665 characters (about 100 words). ROUGE-1 and ROUGE-2 are 0.329 and 0.073 respectively. It confirms that the performance of our approach is sensitive to the length of the summary.

Sum_length	Rouge-1	Rouge-2	Rouge-L
50	0.241	0.036	0.205
100	0.309	0.085	0.277
200	0.396	0.116	0.358
400	**0.423**	**0.118**	**0.402**

Table 7. ROUGE evaluation results for different summary length

7 Conclusions and Future Work

We explore surface, content, event, relevance features and their combinations for extractive summarization with supervised learning approach. Experiments show that the combination of surface, content and relevance features perform best. The highest ROUGE-1, ROUGE-2 scores are 0.396 and 0.116 respectively. The Rouge-1 score of manually generated summaries is 0.422. This shows the ROUGE performance of our supervised learning approach is comparable to that of manually generated summaries. The ROUGE-1 scores of extractive summarization based on centroid, signature word, high frequency word and event individually are 0.319, 0.356, 0.371 and 0.374 respectively. It can be seen that our summarization approach based on combination of features improves the performance obviously.

Although the results of supervised learning approach are encouraging, it required much labeled data. To reduce labeling cost, we apply co-training to combine labeled and unlabeled data. Experiments show that compare with supervised learning, semi-supervised learning approach saves half of the labeling cost and maintains comparable performance (0.366 vs 0.396). We also find that our extractive summarization is sensitive to length of the summary. When the length is extended, the ROUGE scores of same summarization method are improved. In the future, we plan to investigate sentence compression to improve performance of our summarization approaches on short summaries.

Acknowledgement

The research described in this paper is partially supported by Research Grants Council of Hong Kong (RGC: PolyU5217/07E), CUHK Strategic Grant Scheme (No: 4410001) and Direct Grant Scheme (No: 2050417).

References

Regina Barzilay, and Michael Elhadad. 1997. Using lexical chains for text summarization. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics Workshop on Intelligent Scalable Text Summarization*, pages 10-17.

Avrim Blum and Tom Mitchell. 1998. Combining labeled and unlabeled data with co-training. In *Proceedings of the 11th Annual Conference on Computational Learning Theory*, pages 92-100.

Hamish Cunningham, Diana Maynard, Kalina Bontcheva, Valentin Tablan. 2002. GATE: a framework and graphical development environment for robust NLP tools and applications. In *Proceedings of the 40th Annual Meeting of the Association for computational Linguistics*.

Francois Denis and Remi Gilleron. 2003. Text classification and co-training from positive and unlabeled examples. In *Proceedings of the 20th International Conference on Machine Learning Workshop: the Continuum from Labeled Data to Unlabeled Data in Machine Learning and Data Mining*.

Gunes Erkan and Dragomir R. Radev. 2004. LexPageRank: prestige in multi-document text summarization. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 365-371.

Elena Filatova and Vasileios Hatzivassiloglou. Event-based extractive summarization. 2004. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics Workshop*, pages 104-111.

Gerald Francis DeJong. 1978. Fast skimming of news stories: the FRUMP system. Ph.D. thesis, Yale University.

Wenjie Li, Guihong Cao, Kam-Fai Wong and Chunfa Yuan. 2004. Applying machine learning to Chinese temporal relation resolution. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics*, pages 583-589.

Wenjie Li, Wei Xu, Mingli Wu, Chunfa Yuan, Qin Lu. 2006. Extractive summarization using inter- and intra- event relevance. In *proceedings of Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pages 369-376.

Chin-Yew Lin; Eduard Hovy. 2000. The automated acquisition of topic signatures for text summarization. In *Proceedings of the 18th International Conference on Computational Linguistics*, pages 495-501.

Chin-Yew Lin and Eduard Hovy. 2003. Automatic evaluation of summaries using n-gram co-occurrence statistics. In *Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, Edmonton, Canada.

Daniel Marcu. 1997. The rhetorical parsing of natural language texts. In *Proceedings of the 35th Annual Meeting of the Association for computational Linguistics*, pages 96-103.

Christoph Muller, Stefan Rapp and Michael Strube. 2001. Applying co-training to reference resolution. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*.

Ani Nenkova, Lucy Vanderwende and Kathleen McKeown. 2006. A compositional context sensitive multi-document summarizer: exploring the factors that influence summarization. In *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*.

David Pierce and Claire Cardie. 2001. Limitations of co-training for natural language learning from large datasets. In *Proceedings of the 2001 Conference on Empirical Methods in Natural Language Processing*, pages 1-9.

Dragomir R. Radev, Timothy Allison, et al. 2004. MEAD - a platform for multidocument multilingual text summarization. In *Proceedings of 4th International Conference on Language Resources and Evaluation.*

Anoop Sarkar. 2001. Applying co-training methods to statistical parsing. In *Proceedings of 2nd Meeting of the North American Chapter of the Association for Computational Linguistics on Language Technologies.*

Mingli Wu. 2006. Investigations on event-based summarization. In *proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics Student Research Workshop*, pages 37-42.

Mingli Wu, Wenjie Li, Furu Wei, Qin Lu and Kam-Fai Wong. 2007. Exploiting surface, content and relevance features for learning-based extractive summarization. In *Proceedings of 2007 IEEE International Conference on Natural Language Processing and Knowledge Engineering*.

Ting-Fan Wu, Chih-Jen Lin and Ruby C. Weng. 2004. Probability estimates for multi-class classification by pairwise coupling. *Journal of Machine Learning Research*, 5:975-1005.

Masaharu Yoshioka and Makoto Haraguchi. 2004. Multiple news articles summarization based on event reference information. In *Working Notes of the Fourth NTCIR Workshop Meeting*, National Institute of Informatics.