The 22nd International Conference on the Computer Processing of Oriental Languages
(ICCPOL 2009)

CONFERENCE PROGRAM

The Hong Kong Polytechnic University,
Hong Kong

March 26-27, 2009

TABLE OF CONTENTS

CONFERENCE SCHEDULE

(*Page numbers for Abstracts listed on the right*)

Day 1: March 26, 2009			
08:30 – 09:20	**Registration** (Outside of Room AG710)		
09:20 – 09:30	**Open Ceremony** (Room AG710) Welcome Speech by *Kam-Fai Wong*		
09:30 – 10:20	**Keynote Speech 1** (Room AG710) Chair: *Wenjie Li* Information Processing by Human and by Computer: The Case of Language *Professor William S. Y. Wang, The Chinese University of Hong Kong*		
10:20 – 10:50	**Tea Break** (Room AG712)		
10:50 – 12:30	**Session 1: Natural Langue Processing** (Room AG710) Chair: *Zhu Jingbo*		
	10:50-11:15	NLP-01: Fast Semantic Role Labeling for Chinese Based on Semantic Chunking *Weiwei Ding, Baobao Chang*	9
	11:15-11:40	NLP-02: Processing of Korean Natural Language Query Using Local Grammar *Tae-Gil Noh, Yong-Jin Han, Seong-Bae Park, Se-Young Park*	9
	11:40-12:05	NLP-03: A Probabilistic Graphical Model for Recognizing NP Chunks in Texts *Minhua Huang, Robert M. Haralick*	9
	12:05-12:30	NLP-04: CRF Models for Tamil Part of Speech and Chunking *S.Lakshmana Pandian, T.V Geetha*	10
12:30 – 13:30	**Lunch** (On Campus, See Logistics Arrangement)		
13:30 – 14:45	**Session 2: Machine Learning and Web Mining for Natural Langue Processing** (Room AG710) Chair: *Dequan Zheng*		
	13:30-13:55	MLWM-01: A Simple and Efficient Model Pruning Method for Conditional Random Fields *Hai Zhao, Chunyu Kit*	10
	13:55-14:20	MLWM-02: A Density-based Re-ranking Technique for Active Learning for Data Annotations *Jingbo Zhu, Huizhen Wang, Benjamin K Tsou*	10
	14:20-14:45	MLWM-03: Automatic Acquisition of Attributes for Ontology Construction *Gaoying Cui, Qin Lu, Wenjie Li, Yirong Chen*	10
14:45 – 16:05	**Tea Break** and **Poster Session** (Room AG712)		
16:05 – 17:45	**Session 3: Machine Translation** (Room AG710) Chair: *Kevin Knight*		
	16:05-16:30	MT-01: Lexicalized Syntactic Reordering Framework for Word Alignment and Machine Translation *Chung-Chi Huang, Wei-Teh Chen, Jason S. Chang*	11
	16:30-16:55	MT-02: Validity of an Automatic Evaluation of Machine Translation Using a Word-alignment-based Classifier *Katsunori Kotani, Takehiko Yoshimi, Takeshi Kutsumi, Ichiko Sata*	11
	16:55-17:20	MT-03: Harvesting Regional Transliteration Variants with Guided Search *Jin-Shea Kuo, Haizhou Li, Chih-Lung Lin*	11
	17:20-17:45	MT-04: Found in Translation: Conveying Subjectivity of a Lexicon of One Language into Another Using a Bilingual Dictionary and a Link Analysis Algorithm *Jungi Kim, Hun-Young Jung, Sang-Hyeob Nam, Yeha Lee, Jong-Hyeok Lee*	12
18:30	**Banquet** (See Logistics Arrangement)		

Day 2: March 27, 2009			
09:00 – 09:30	**Registration** (Outside of Room AG710**)**		
09:30 – 10:20	**Keynote Speech 2** (Room AG710) Chair: *Chu-Ren Huang* KOTONOHA: A Corpus Compilation Initiative at the National Institute for Japanese Language *Professor Kikuo Maekawa, The National Institute for Japanese Language*		
10:20 – 10:50	**Tea Break** (Outside of Room AG710)		
10:50 – 12:30	**Session 4: Information Extraction** (Room AG710) Chair: *Jin-Shea Kuo*		
	10:50-11:15	IE-01: A Novel Composite Kernel Approach to Chinese Entity Relation Extraction *Ji Zhang, You Ouyang, Wenjie Li, Yuexian Hou*	12
	11:15-11:40	IE-02: A Novel Method of Automobiles' Chinese Nickname Recognition *Cheng Wang, Wenyuan Yu, Wenxin Li, Zhuoqun Xu*	12
	11:40-12:05	IE-03: Research on Automatic Chinese Multi-word Term Extraction Based on Term Component *Wei Kang, Zhifang Sui*	13
	12:05-12:30	IE-04: Speech Synthesis for Error Training Models in CALL *Xin Zhang, Qin Lu, Jiping Wan, Guangguang Ma, Tin Shing Chiu, Weiping Ye, Wenli Zhou, Qiao Li*	13
12:30 – 13:30	**Lunch** (On Campus, See Logistics Arrangement)		
13:30 – 15:10	**Session 5: Document Summarization** (Room AG710) Chair: *Takehito Utsuro*		
	13:30-13:55	SUMM-01: An Extractive Text Summarizer Based on Significant Words *Xiaoyue Liu, Jonathan Webster, Chunyu Kit*	13
	13:55-14:20	SUMM-02: Query-Oriented Summarization Based on Neighborhood Graph Model *Furu Wei, Yanxiang He, Wenjie Li, Lei Huang*	13
	14:20-14:45	SUMM-03: Using Proximity in Query Focused Multi-document Extractive Summarization *Sujian Li, Yu Zhang, Wei Wang, Chen Wang*	14
	14:45-15:10	SUMM-04: Learning Similarity Functions in Graph-based Document Summarization *You Ouyang, Wenjie Li, Furu Wei, Qin Lu*	14
15:10 – 15:40	**Tea Break** (Outside of Room AG710)		
15:40 – 17:20	**Session 6: Sentiment Classification and Other Topics** (Room AG710) Chair: *Ruifeng Xu*		
	15:40-16:05	SCOT-01: Partially Supervised Phrase-Level Sentiment Classification *Sang-Hyob Nam, Seung-Hoon Na, Jungi Kim, Yeha Lee, Jong-Hyeok Lee*	14
	16:05-16:30	SCOT-02: Extracting Domain-Dependent Semantic Orientations of Latent Variables for Sentiment Classification *Yeha Lee, Jungi Kim, Jong-Hyeok Lee*	15
	16:30-16:55	SCOT-03: Mining Cross-Lingual/Cross-Cultural Differences in Concerns and Opinions in Blogs *Hiroyuki Nakasaki, Mariko Kawaba, Takehito Utsuro, Tomohiro Fukuhara*	15
	16:55-17:20	SCOT-04: Probabilistic Methods for a Japanese Syllable Cipher *Sujith Ravi, Kevin Knight*	15
17:30 – 17:40	**Closing Ceremony** (Room AG710) Closing Speech by *Qin Lu*		

KEYNOTE SPEECH BY PROF. WILLIAM S.W. WANG

Information Processing by Human and by Computer: The Case of Language

Professor William S. Y. Wang

The Chinese University of Hong Kong

During the last days before his death in 1957, John Von Neumann worked on a lecture comparing the computer and the brain. The publication has been recently reprinted in 2000, with a very helpful update by Paul and Patricia Churchland. I will follow up on the comparison, focusing mainly on the processing of linguistic information. The computer is able to outperform humans in closed tasks, such as the game of chess, because of its superiority in accuracy and speed. On the other hand, tasks such as the recognition of speech sounds, and the translation from one language to another, continue to elude the reach of the computer. These tasks require access to open-ended bodies of real world knowledge, much of which remain unsystematized. With recent advances in the cognitive neurosciences, we are just beginning to understand how such information is organized and accessed in the brain.

KEYNOTE SPEECH BY PROF. KIKUO MAEKAWA

KOTONOHA: A Corpus Compilation Initiative at the National Institute for Japanese Language

Professor Kikuo Maekawa

The National Institute for Japanese Language

Today's natural language and speech processing technology depends heavily upon statistical approach; hence there is need for large-scale language corpora. As long as the Japanese language is concerned, it was not linguists but engineers that played a leading role for the development of language and speech corpora in the early stage of corpus-related studies.

In the last half of the 1990s, however, linguists became more and more interested in the corpus-bases analysis of the Japanese language. But it turned out quickly that the characteristics of corpora needed in engineering and linguistics did not necessarily coincide. Most obvious difference consisted in the temporal coverage of corpora; while engineers dealt with the data of contemporary language, linguists often had a need for the data of the past language. Also obvious was the difference in the size and quality of data: in many cases, engineers put more emphasis on the size rather than the quality of data. Lastly, it was also the case that, generally speaking, linguists were interested in much wider range of language varieties than engineers.

The aim of the KOTONOHA initiative of the National Institute for Japanese Language (NIJL) consists in providing a series of corpora that cover the whole range of the Japanese language. KOTONOHA (ancient Japanese meaning 'language') is a cover term for a series of component corpora covering both written and spoken, and, contemporary as well as historical varieties. The component corpora are designed primarily for linguistic research, but the possibility of engineering application is not excluded from the corpus design, especially when a corpus is concerned with contemporary spoken Japanese.

So far, there are 2 corpora of KOTONOHA that have already been made publicly available. The Taiyô Corpus deals with the written Japanese of at around the boundary of the 19th and 20th centuries; it is the time when the grammar of the written Japanese changed drastically. The Corpus of Spontaneous Japanese (CSJ) is a spoken language corpus designed for the statistical learning of the language- and acoustic-models for the next generation automatic speech recognition system that is capable of handling more-or-less spontaneous speech. Although the CSJ was built primarily for engineering application, a subset of the corpus was richly annotated for the phonetic and linguistic study of spontaneous speech. The annotation included X-JToBI labels for segmental and prosodic features, clause boundary labels, and, the labels of topic boundary.

Since 2006, the corpus compilation group of the NIJL devotes concentrated efforts to the compilation of the Balanced Corpus of Contemporary Written Japanese (BCCWJ). It is a long-awaited balanced corpus of contemporary Japanese and probably the most important corpus of the whole KOTONOHA initiative. Among the 100 million words samples of the BCCWJ, about 35% is chosen by means of random sampling from the populations covering the books, magazines, and newspapers published in the years 2001-2005. In addition, 30% of the samples are randomly chosen from the population of the books registered in the public libraries of Tokyo Metropolis and published during the years 1985-2005. And, the resulting 35% is devoted to various mini corpora including the texts of governmental white papers, minute of the National Diet, textbooks of elementary and secondary education, bestselling books, the Internet texts of blog and bulletin board, and so forth.

Although the compilation of the BCCWJ as a whole is not an easy task, the biggest problem lies in the clearance of copyright protected materials. For example, as long as the book samples are concerned, more than 10,000 permissions from the copyright holders have been obtained during the last 30 months, but they cover only 40% of the totality of 24,000 book texts that have been sampled so far. This is not because the copyright holders were unwilling to give permission. More than 90% of copyright holders gave us permission, once we could make contact with them. The problem lies in the fact that it is often impossible to make contact with the copyright holders.

Though we will continue working hard to obtain permissions from the copyright holders, an amendment of Japanese copyright law by which ‘fair-use’ of copyright protected materials becomes possible is strongly needed. Fortunately, the Copyright Division of the Agency of Cultural Affairs is now preparing a bill for such amendment. At the earliest, the deliberation of the amendment bill will be started sometime in 2009.

Anyway, corpus materials corresponding to about 70 million words have already been sampled and stored in a server, and about 40 million words copyright-cleared texts are available for full-text search in a demonstration web site (http://www.kotonoha.gr.jp/demo/). The BCCWJ will be publicly available in 2011.

ABSTRACTS OF ORAL PRESENTATIONS

NLP-01: Fast Semantic Role Labeling for Chinese Based on Semantic Chunking

Weiwei Ding, Baobao Chang

Recently, with the development of Chinese semantically annotated corpora, e.g. the Chinese Proposition Bank, the Chinese semantic role labeling (SRL) has been boosted. However, the Chinese SRL researchers now focus on the transplant of existing statistical machine learning methods which have been proven to be effective on English. In this paper, we have established a semantic chunking based method which is different from the traditional ones. Semantic chunking is named because of its similarity with syntactic chunking. The difference is that semantic chunking is used to identify the semantic chunks, i.e. the semantic roles. Based on semantic chunking, the process of SRL is changed from "parsing – semantic role identification – semantic role classification", to "semantic chunk identification – semantic chunk classification". With the elimination of the parsing stage, the SRL task can get rid of the dependency on parsing, which is the bottleneck both of speed and precision. The experiments have shown that the semantic chunking based method outperforms previously best-reported results on Chinese SRL and saves a large amount of time.

NLP-02: Processing of Korean Natural Language Query Using Local Grammar

Tae-Gil Noh, Yong-Jin Han, Seong-Bae Park, Se-Young Park

For casual web users, a natural language is more accessible than formal query languages. However, understanding of a natural language query is not trivial for computer systems. This paper proposes a method to parse and understand Korean natural language queries with local grammars. A local grammar is a formalism that can model syntactic structures and synonymous phrases. With local grammars, the system directly extracts user's intentions from natural language queries. With 163 hand-crafted local grammar graphs, the system could attain a good level of accuracy and meaningful coverage over IT company/people domain.

NLP-03: A Probabilistic Graphical Model for Recognizing NP Chunks in Texts

Minhua Huang, Robert M. Haralick

We present a probabilistic graphical model for identifying noun phrase patterns in texts. This model is derived from mathematical processes under two reasonable conditional independence assumptions with different perspectives compared with other graphical models, such as CRFs or MEMMs. Empirical results shown our model is effective. Experiments on WSJ data from the Penn Treebank, our method achieves an average of precision 97.7% and an average of recall 98.7%. Further experiments on the CoNLL-2000 shared task data set show our method achieves the best performance compared to competing methods that other researchers have published on this data set. Our average precision is 95.15% and an average recall is 96.05%.

NLP-04: CRF Models for Tamil Part of Speech and Chunking

S.Lakshmana Pandian, T.V Geetha

Conditional random fields (CRFs) is a framework for building probabilistic models to segment and label sequence data. CRFs offer several advantages over hidden Markov models (HMMs) and stochastic grammars for such tasks, including the ability to relax strong independence assumptions made in those models. CRFs also avoid a fundamental limitation of maximum entropy Markov models (MEMMs) and other discriminative Markov models based on directed graphical models, which can be biased towards states with few successor states. In this paper we propose the Language Models developed for Part Of Speech (POS) tagging and chunking using CRFs for Tamil. The Language models are designed based on morphological information. The CRF based POS tagger has an accuracy of about 89.18%, for Tamil and the chunking process performs at an accuracy of 84.25% for the same language.

MLWM-01: A Simple and Efficient Model Pruning Method for Conditional Random Fields

Hai Zhao, Chunyu Kit

Conditional random fields (CRFs) have been quite successful in various machine learning tasks. However, as larger and larger data become acceptable for the current computational machines, trained CRFs Models for a real application quickly inflate. Recently, researchers often have to use models with tens of millions features. This paper considers pruning an existing CRFs model for storage reduction and decoding speedup. We propose a simple but efficient rank metric for feature group rather than features that previous work usually focus on. A series of experiments in two typical labeling tasks, word segmentation and named entity recognition for Chinese, are carried out to check the effectiveness of the proposed method. The results are quite positive and show that CRFs models are highly redundant, even using carefully selected label set and feature templates.

MLWM-02: A Density-based Re-ranking Technique for Active Learning for Data Annotations

Jingbo Zhu, Huizhen Wang, Benjamin K Tsou

One of the popular techniques of active learning for data annotations is uncertainty sampling, however, which often presents problems when outliers are selected. To solve this problem, this paper proposes a density-based re-ranking technique, in which a density measure is adopted to determine whether an unlabeled example is an outlier. The motivation of this study is to prefer not only the most informative example in terms of uncertainty measure, but also the most representative example in terms of density measure. Experimental results of active learning for word sense disambiguation and text classification tasks using six real-world evaluation data sets show that our proposed density-based re-ranking technique can improve uncertainty sampling.

MLWM-03: Automatic Acquisition of Attributes for Ontology Construction

Gaoying Cui, Qin Lu, Wenjie Li, Yirong Chen

An ontology can be seen as an organized structure of concepts according to their relations. A concept is associated with a set of attributes that themselves are also concepts in the ontology. Consequently, ontology construction is the acquisition of concepts and their associated attributes through relations. Manual ontology construction is time-consuming and difficult to maintain. Corpus-based ontology construction methods must be able to distinguish concepts themselves from concept instances. In this paper, a novel and simple method is proposed for automatically identifying concept attributes through the use of Wikipedia as the corpus. The built-in {{Infobox}} in Wiki is used to acquire concept attributes and identify semantic types of the attributes. Two simple induction rules are applied to improve the performance. Experimental results show precisions of 92.5% for attribute acquisition and 80% for attribute type identification. This is a very promising result for automatic ontology construction.

MT-01: Lexicalized Syntactic Reordering Framework for Word Alignment and Machine Translation

Chung-Chi Huang, Wei-Teh Chen, Jason S. Chang

We propose a lexicalized syntactic reordering framework for cross-language word aligning and translating researches. In this framework, we first flatten hierarchical source-language parse trees into syntactically-motivated linear string representations, which can easily be input to many feature-like probabilistic models. During model training, these string representations accompanied with target-language word alignment information are leveraged to learn systematic similarities and differences in languages' grammars. At runtime, syntactic constituents of source-language parse trees will be reordered according to automatically acquired lexicalized reordering rules in previous step, to closer match word orientations of the target language. Empirical results show that, as a preprocessing component, bilingual word aligning and translating tasks benefit from our reordering methodology.

MT-02: Validity of an Automatic Evaluation of Machine Translation Using a Word-alignment-based Classifier

Katsunori Kotani, Takehiko Yoshimi, Takeshi Kutsumi, Ichiko Sata

Because human evaluation of machine translation is extensive but expensive, we often use automatic evaluation in developing a machine translation system. From viewpoint of evaluation cost, there are two types of evaluation methods: one uses (multiple) reference translation, e.g., METEOR, and the other classifies machine translation either into machine-like or human-like translation based on translation properties, i.e., a classification-based method. Previous studies showed that classification-based methods could perform evaluation properly. These studies constructed a classifier by learning linguistic properties of translation such as length of a sentence, syntactic complexity, and literal translation, and their classifiers marked high classification accuracy. These previous studies, however, have not examined whether their classification accuracy could present translation quality. Hence, we investigated whether classification accuracy depends on translation quality. The experiment results showed that our method could correctly distinguish the degrees of translation quality.

MT-03: Harvesting Regional Transliteration Variants with Guided Search

Jin-Shea Kuo, Haizhou Li, Chih-Lung Lin

This paper proposes a method to harvest regional transliteration variants with guided search. We first study how to incorporate transliteration knowledge into query formulation so as to significantly increase the chance of desired transliteration returns. Then, we study a cross-training algorithm, which explores valuable information across different regional corpora for the learning of transliteration models to in turn improve the overall extraction performance. The experimental results show that the proposed method not only effectively harvests a lexicon of regional transliteration variants but also mitigates the need of manual data labeling for transliteration modeling. We also conduct an inquiry into the underlying characteristics of regional transliterations that motivate the cross-training algorithm.

MT-04: Found in Translation: Conveying Subjectivity of a Lexicon of One Language into Another Using a Bilingual Dictionary and a Link Analysis Algorithm

Jungi Kim, Hun-Young Jung, Sang-Hyeob Nam, Yeha Lee, Jong-Hyeok Lee

This paper proposes a method that automatically creates a subjectivity lexicon in a new language using a subjectivity lexicon in a resource–rich language with only a bilingual dictionary. We resolve some of the difficulties in selecting appropriate senses when translating lexicon, and present a framework that sequentially applies an iterative link analysis algorithm to enhance the quality of lexicons of both the source and target languages. The experimental results have empirically shown to improve the subjectivity lexicon in the source language as well as create a good quality lexicon in a new language.

IE-01: A Novel Composite Kernel Approach to Chinese Entity Relation Extraction

Ji Zhang, You Ouyang, Wenjie Li, Yuexian Hou

Relation extraction is the task of finding semantic relations between two entities from the text. In this paper, we propose a novel composite kernel for Chinese relation extraction. The composite kernel is defined as the combination of two independent kernels. One is the entity kernel built upon the non-content-related features. The other is the string semantic similarity kernel concerning the content information. Three combinations, namely linear combination, semi-polynomial combination and polynomial combination are investigated. When evaluated on the ACE 2005 Chinese data set, the results show that the proposed approach is effective.

IE-02: A Novel Method of Automobiles' Chinese Nickname Recognition

Cheng Wang, Wenyuan Yu, Wenxin Li, Zhuoqun Xu

Nowadays, we have noticed that the free writing style becomes more and more popular. People tend to use nicknames to replace the original names. However, the traditional named entity recognition does not perform well on the nickname recognition problem. Thus, we chose the automobile domain and accomplished a whole process of Chinese automobiles' nickname recognition. This paper discusses a new method to tackle the problem of automobile's nickname recognition in Chinese text. First we have given the nicknames a typical definition. Then we have used methods of machine learning to acquire the probabilities of transition and emission based on our training set. Finally the nicknames are identified through maximum matching on the optimal state sequence. The result revealed that our method can achieve competitive performance in nickname recognition. We got precision 95.2%; recall 91.5% and F-measure 0.9331 on our passages test set. The method will contribute to build a database of nicknames, and could be used in data mining and search engines on automobile domain, etc.

IE-03: Research on Automatic Chinese Multi-word Term Extraction Based on Term Component
Wei Kang, Zhifang Sui

This paper presents an automatic Chinese multi-word term extraction method based on the unithood and the termhood measure. The unithood of the candidate term is measured by the strength of inner unity and marginal variety. Term component is taken into account to estimate the termhood. Inspired by the economical law of term generating, we propose two measures of a candidate term to be a true term: the first measure is based on domain speciality of term, and the second one is based on the similarity between a candidate and a template that contains structured information of terms. Experiments on I.T. domain and Medicine domain show that our method is effective and portable in different domains.

IE-04: Speech Synthesis for Error Training Models in CALL
Xin Zhang, Qin Lu, Jiping Wan, Guangguang Ma, Tin Shing Chiu, Weiping Ye, Wenli Zhou, Qiao Li

A computer assisted pronunciation teaching system (CAPT) is a fundamental component in a computer assisted language learning system (CALL). A speech recognition based CAPT system often requires a large amount of speech data to train the incorrect phone models in its speech recognizer. But collecting incorrectly pronounced speech data is a labor intensive and costly work. This paper reports an effort on training the incorrect phone models by making use of synthesized speech data. A special formant speech synthesizer is designed to filter the correctly pronounced phones into incorrect phones by modifying the formant frequencies. In a Chinese Putonghua CALL system for native Cantonese speakers to learn Mandarin, a small experimental CAPT system is built with a synthetic speech data trained recognizer. Evaluation shows that a CAPT system using synthesized data can perform as good as or even better than that using real data provided that the size of the synthetic data are large enough.

SUMM-01: An Extractive Text Summarizer Based on Significant Words
Xiaoyue Liu, Jonathan Webster, Chunyu Kit

Document summarization can be viewed as a reductive distilling of source text through content condensation, while words with high quantities of information are believed to carry more content and thereby importance. In this paper, we propose a new quantification measure for word significance used in natural language processing (NLP) tasks, and successfully apply it to an extractive text summarization approach. In a query-based summarization setting, the correlation between user queries and sentences to be scored is established from both the micro (i.e. at the word level) and the macro (i.e. at the sentence level) perspectives, resulting in an effective ranking formula. The experiments, both on a generic single document summarization evaluation, and on a query-based multi-document evaluation, verify the effectiveness of the proposed measures and show that the proposed approach achieves a state-of-the-art performance.

SUMM-02: Query-Oriented Summarization Based on Neighborhood Graph Model
Furu Wei, Yanxiang He, Wenjie Li, Lei Huang

In this paper, we investigate how to combine the link-aware and link-free information in sentence ranking for query-oriented summarization. Although the link structure has been emphasized in the existing graph-based summarization models, there is lack of pertinent analysis on how to use the links. By contrasting the text graph with the web graph, we propose to evaluate significance of sentences based on neighborhood graph model. Taking the advantage of the link information provided on the graph, each sentence is evaluated according to its own value as well as the cumulative impacts from its neighbors. For a task like query-oriented summarization, it is critical to explore how to reflect the influence of the query. To better incorporate query information into the model, we further design a query-sensitive similarity measure to estimate the association between a pair of sentences. When evaluated on DUC 2005 dataset, the results of the pro-posed approach are promising.

SUMM-03: Using Proximity in Query Focused Multi-document Extractive Summarization

Sujian Li, Yu Zhang, Wei Wang, Chen Wang

The query focused multi-document summarization tasks usually tend to answer the queries in the summary. In this paper, we suggest introducing an effective feature which can represent the relation of key terms in the query. Here, we adopt the feature of term proximity commonly used in the field of information retrieval, which has improved the retrieval performance according to the relative position of terms. To resolve the problem of data sparseness and to represent the proximity in the semantic level, concept expansion is conducted based on WordNet. By leveraging the term importance, the proximity feature is further improved and weighted according to the inverse term frequency of terms. The experimental results show that our proposed feature can contribute to improving the summarization performance.

SUMM-04: Learning Similarity Functions in Graph-based Document Summarization

You Ouyang, Wenjie Li, Furu Wei, Qin Lu

Graph-based models have been extensively explored in document summarization in recent years. Compared with traditional feature-based models, graph-based models incorporate interrelated information into the ranking process. Thus, potentially they can do a better job in retrieving the important contents from documents. In this paper, we investigate the problem of how to measure sentence similarity which is a crucial issue in graph-based summarization models but in our belief has not been well defined in the past. We propose a supervised learning approach that brings together multiple similarity measures and makes use of human-generated summaries to guide the combination process. Therefore, it can be expected to provide more accurate estimation than a single cosine similarity measure. Experiments conducted on the DUC2005 and DUC2006 data sets show that the proposed learning approach is successful in measuring similarity. Its competitiveness and adaptability are also demonstrated.

SCOT-01: Partially Supervised Phrase-Level Sentiment Classification

Sang-Hyob Nam, Seung-Hoon Na, Jungi Kim, Yeha Lee, Jong-Hyeok Lee

This paper presents a new partially supervised approach to phrase-level sentiment analysis that first automatically constructs a polarity-tagged corpus and then learns sequential sentiment tag from the corpus. This approach uses only sentiment sentences which are readily available on the Internet and does not use a polarity-tagged corpus which is hard to construct manually. With this approach, the system is able to automatically classify phrase-level sentiment. The result shows that a system can learn sentiment expressions without a polarity-tagged corpus.

SCOT-02: Extracting Domain-Dependent Semantic Orientations of Latent Variables for Sentiment Classification

Yeha Lee, Jungi Kim, Jong-Hyeok Lee

Sentiment analysis of weblogs is a challenging problem. Most previous work utilized semantic orientations of words or phrases to classify sentiments of weblogs. The problem with this approach is that semantic orientations of words or phrases are investigated without considering the domain of weblogs. Weblogs contain the author's various opinions about multifaceted topics. Therefore, we have to treat a semantic orientation domain-dependently. In this paper, we present an unsupervised learning model based on aspect model to classify sentiments of weblogs. Our model utilizes domain-dependent semantic orientations of latent variables instead of words or phrases, and uses them to classify sentiments of weblogs. Experiments on several domains confirm that our model assigns domain-dependent semantic orientations to latent variables correctly, and classifies sentiments of weblogs effectively.

SCOT-03: Mining Cross-Lingual/Cross-Cultural Differences in Concerns and Opinions in Blogs

Hiroyuki Nakasaki, Mariko Kawaba, Takehito Utsuro, Tomohiro Fukuhara

The goal of this paper is to cross-lingually analyze multilingual blogs collected with a topic keyword. The framework of collecting multilingual blogs with a topic keyword is designed as the blog feed retrieval procedure. Mulitlingual queries for retrieving blog feeds are created from Wikipedia entries. Finally, we cross-lingually and cross-culturally compare less well known facts and opinions that are closely related to a given topic. Preliminary evaluation results support the effectiveness of the proposed framework.

SCOT-04: Probabilistic Methods for a Japanese Syllable Cipher

Sujith Ravi, Kevin Knight

This paper attacks a Japanese syllable-substitution cipher. We use a probabilistic, noisy-channel framework, exploiting various Japanese language models to drive the decipherment. We describe several innovations, including a new objective function for searching for the highest-scoring decipherment. We include empirical studies of the relevant phenomena, and we give improved decipherment accuracy rates.

ABSTRACTS OF POSTERS

POSTER-01: Acquiring Verb Subcategorization Frames in Bengali from Corpora

Dipankar Das, Asif Ekbal, and Sivaji Bandyopadhyay

Subcategorization frames acquisition of a phrase can be described as a mechanism to extract different types of relevant arguments that are associated with that phrase in a sentence. This paper presents the acquisition of different subcategory frames for a specific Bengali verb that has been identified from POS tagged and chunked data prepared from raw Bengali news corpus. Syntax plays the main role in the acquisition process and not the semantics like thematic roles. The output frames of the verb have been compared with the frames of its English verb that has been identified using bilingual lexicon. The frames for the English verb have been extracted using Verbnet. This system has demonstrated precision and recall values of 85.21% and 83.94% respectively on a test set of 1500 sentences.

POSTER-02 An Integrated Approach for Concept Learning and Relation Extraction

Qingliang Zhao and Zhifang Sui

Concept learning and hierarchical relations extraction are core tasks of ontology automatic construction. In the current research, the two tasks are carried out separately, which separates the natural association between them. This paper proposes an integrated approach to do the two tasks together. The attribute values of concepts are used to evaluate the extracted hierarchical relations. On the other hand, the extracted hierarchical relations are used to expand and evaluate the attribute values of concepts. Since the interaction is based on the inaccurate result that extracted automatically, we introduce the weight of intermediate results of both tasks into the iteration to ensure the accuracy of results. Experiments have been carried out to compare the integrated approach with the separated ones for concept learning and hierarchical relations. Our experiments show performance improvements in both tasks.

POSTER-03 An Investigation of an Interontologia: Comparison of the Thousand-Character Text and Roget's Thesaurus

Sang-Rak Kim, Jae-Gun Yang, and Jae-Hak J. Bae

The present study presents the lexical category analysis of the Thousand-Character Text and Roget's Thesaurus. Through preprocessing, the Thousand-Character Text and Roget's Thesaurus have been built into databases. In addition, for easier analysis and more efficient research, we have developed a system to search Roget's Thesaurus for the categories corresponding to Chinese characters in the Thousand-Character Text. According to the results of this study, most of the 39 sections of Roget's Thesaurus except the 'Creative Thought' section were relevant to Chinese characters in the Thousand-Character Text. Three sections 'Space in General', 'Dimensions' and 'Matter in General' have higher mapping rate. The correlation coefficient is also around 0.94, showing high category relevancy on the section level between the Thousand-Character Text and Roget's Thesaurus.

POSTER-04 Constructing Parallel Corpus from Movie Subtitles

Han Xiao and Xiaojie Wang

This paper describes a methodology for constructing aligned German-Chinese corpora from movie subtitles. The corpora will be used to train a special machine translation system with intention to automatically translate the subtitles between German and Chinese. Since the common length-based algorithm for alignment shows weakness on short spoken sentences, especially on those from different language families, this paper studies to use dynamic programming based on time-shift information in subtitles, and extends it with statistical lexical cues to align the subtitle. In our experiment with around 4,000 Chinese and German sentences, the proposed alignment approach yields 83.8% precision. Furthermore, it is unrelated to languages, and leads to a general method of parallel corpora building between different language families.

POSTER-05 Dialogue Strategies to Overcome Speech Recognition Errors in Form-Filling Dialogue

Sangwoo Kang, Songwook Lee, and Jungyun Seo

In a spoken dialogue system, the speech recognition performance accounts for the largest part of the overall system performance. Yet spontaneous speech recognition has an unstable performance. The proposed postprocessing method solves this problem. The state of a legacy DB can be used as an important factor for recognizing a user’s intention because form-filling dialogues tend to depend on the legacy DB. Our system uses the legacy DB and ASR result to infer the user’s intention, and the validity of the current user’s intention is verified using the inferred user’s intention. With a plan-based dialogue model, the proposed system corrected 27% of the incomplete tasks, and achieved an 89% overall task completion rate.

POSTER-06 Document Clustering Description Extraction and Its Application

Chengzhi Zhang, Huilin Wang, Yao Liu, and Hongjiao Xu

Document clustering description is a problem of labeling the clustering results of document collection clustering. It can help users determine whether one of the clusters is relevant to their information requirements or not. To resolve the problem of the weak readability of document clustering results, a method of automatic labeling document clusters based on machine learning is put forward. Clustering description extraction in application to topic digital library construction is introduced firstly. Then, the descriptive results of five models are analyzed respectively, and their performances are compared.

POSTER-07 Domain Adaptation for English–Korean MT System: From Patent Domain to IT Web News Domain

Ki-Young Lee, Sung-Kwon Choi, Oh-Woog Kwon, Yoon-Hyung Roh, and Young-Gil Kim

This paper addresses a method to adapt an existing machine translation (MT) system for a newly targeted translation domain. Especially, we give detailed descriptions of customizing a patent specific English-Korean machine translation system for IT web news domain. The proposed method includes the followings: constructing a corpus from documents of IT web news domain, analyzing characteristics of IT web news sentences according to each viewpoint of MT system modules (tagger, parser, transfer) and translation knowledge, and adapting each MT system modules and translation knowledge considering characteristics of IT web news domain. To evaluate our domain adaptation method, we conducted a human evaluation and an automatic evaluation. The experiment showed promising results for diverse sentences extracted from IT Web News documents.

POSTER-08 Event-Based Summarization Using Critical Temporal Event Term Chain

Maofu Liu, Wenjie Li, Xiaolong Zhang, and Ji Zhang

In this paper, we investigate whether temporal relations among event terms can help improve event-based summarization and text cohesion of final summaries. By connecting event terms with happens-before relations, we build a temporal event term graph for source documents. The event terms in the critical temporal event term chain identified from the maximal weakly connected component are used to evaluate the sentences in source documents. The most significant sentences are included in final summaries. Experiments conducted on the DUC 2001 corpus show that event-based summarization using the critical temporal event term chain is able to organize final summaries in a more coherent way and make improvement over the well-known tf*idf-based and PageRank-based summarization approaches.

POSTER-09 Experiment Research on Feature Selection and Learning Method in Keyphrase Extraction

Chen Wang, Sujian Li, and Wei Wang

Keyphrases can provide a brief summary of documents. Keyphrase extraction, defined as automatic selection of important phrases within the body of a document, is important in some fields. Generally the keyphrase extraction process is seen as a classification task, where feature selection and learning model are the key problems. In this paper, different shallow features are surveyed and the commonly used learning methods are compared. The experimental results demonstrate that the detailed survey of shallow features plus a simpler method can more enhance the extraction performance.

POSTER-10 Flattened Syntactical Phrase-Based Translation Model for SMT

Qing Chen and Tianshun Yao

This paper proposed a flattened syntactical phrase-based translation model for Statistical Machine Translation (SMT) learned from bilingual parallel parsed texts. The flattened syntactical phrases are sets of ordered leaf nodes with their father nodes of single syntax trees or forests ignoring the inner structure, containing lexicalized terminals and non-terminals as variable nodes. Constraints over the variable nodes in target side guarantee correct syntactical structures of translations in accordant to the syntactical knowledge learned from parallel texts. The experiments based on Chinese-to-English translation show us a predictable result that our model achieves 1.87% and 4.76% relative improvements, over Pharaoh, the state-of-art phrase-based translation system, and the system of traditional tree-to-tree model based on STSG.

POSTER-11 Korean-Chinese Machine Translation Using Three-Stage Verb Pattern Matching

Chang Hao Yin, Young Ae Seo, and Young-Gil Kim

In this paper, we describe three-stage pattern matching approach and an effective pattern construction process in the Korean-Chinese Machine Translation System for technical documents. We automatically extracted about 100,000 default verb patterns and about 10,000 default ordering patterns from the existing patterns. With the new three-stage approach, additionally using default verb and ordering patterns, the matching hit rate increases 3.8% , comparing with one-stage approach.

POSTER-12: Meta-evaluation of Machine Translation Using Parallel Legal Texts

Billy Tak-Ming Wong and Chunyu Kit

In this paper we report our recent work on the evaluation of a number of popular automatic evaluation metrics for machine translation using parallel legal texts. The evaluation is carried out, following a recognized evaluation protocol, to assess the reliability, the strengths and weaknesses of these evaluation metrics in terms of their correlation with human judgment of translation quality. The evaluation results confirm the reliability of the well-known evaluation metrics, BLEU and NIST for English-to-Chinese translation, and also show that our evaluation metric ATEC outperforms all others for Chinese-to-English translation. We also demonstrate the remarkable impact of different evaluation metrics on the ranking of online machine translation systems for legal translation.

POSTER-13: PKUNEI – A Knowledge–Based Approach for Chinese Product Named Entity Semantic Identification

Wenyuan Yu, Cheng Wang, Wenxin Li, and Zhuoqun Xu

We present the idea of Named Entity Semantic Identification that is identifying the named entity in a knowledge base and give a definition of this idea. Then we introduced PKUNEI - an approach for Chinese product named entity semantic identification. This approach divided the whole process into 2 separate phases: a role-model based NER phase and a query-driven semantic identification phase. We describe the model of NER phase, the automatically building of knowledge base and the implementation of semantic identification phase. The experimental results demonstrate that our approach is effective for the semantic identification task.

POSTER-14: Research on Domain Term Extraction Based on Conditional Random Fields

Dequan Zheng, Tiejun Zhao, and Jing Yang

Domain Term Extraction has an important significance in natural language processing, and it is widely applied in information retrieval, information extraction, data mining, machine translation and other information processing fields. In this paper, an automatic domain term extraction method is proposed based on condition random fields. We treat domain terms extraction as a sequence labeling problem, and terms' distribution characteristics as features of the CRF model. Then we used the CRF tool to train a template for the term extraction. Experimental results showed that the method is simple, with common domains, and good results were achieved. In the open test, the precision rate achieved was 79.63 %, recall rate was 73.54%, and F-measure was 76.46%.

POSTER-15: Text Editing for Lecture Speech Archiving on the Web

Masashi Ito, Tomohiro Ohno, and Shigeki Matsubara

It is very significant in the knowledge society to accumulate spoken documents on the web. However, because of the high redundancy of spontaneous speech, the transcribed text in itself is not readable on an Internet browser, and therefore not suitable as a web document. This paper proposes a technique for converting spoken documents into web documents for the purpose of building a speech archiving system. The technique edits automatically transcribed texts and improves its readability on the browser. The readable text can be generated by applying technology such as paraphrasing, segmentation and structuring to the transcribed texts. An edit experiment using lecture data showed the feasibility of the technique. A prototype system of spoken document archiving was implemented to confirm its effectiveness.

AUTHOR INDEX

ICCPOL 2009 COMMITTEES

Honorary Chairs

Bonnie Dorr	University of Maryland, USA (2008 President, ACL)
Jong Hyeok Lee	POSTECH, Korea (President, COLCS)
Elizabeth D. Liddy	Syracuse University, USA (Chair, ACM SIGIR)
Jun-Ichi Tsujii	University of Tokyo Japan (President, AFNLP)

Conference Chairs

Qin Lu	The Hong Kong Polytechnic University, Hong Kong
Robert Dale	Macquarie University, Australia

Program Chairs

Wenjie Li	The Hong Kong Polytechnic University, Hong Kong
Diego Molla	Macquarie University, Australia

Local Organization Chair

Grace Ngai	The Hong Kong Polytechnic University, Hong Kong

Local Organization Committee Members

Gaoying Cui	The Hong Kong Polytechnic University, Hong Kong
You Ouyang	The Hong Kong Polytechnic University, Hong Kong

Publication Chairs

Chunyu Kit	City University of Hong Kong, Hong Kong
Ruifeng Xu	City University of Hong Kong, Hong Kong

Program Committee Members

(starting with last names)

Tim Baldwin	University of Melbourne, Australia
Sivaji Bandyopadhyay	Jadavpur University, India
Hsin-Hsi Chen	National Taiwan University, Taiwan
Keh-Jiann Chen	Academia Sinica, Taiwan
Key-Sun Choi	Korea Advanced Institute of Science and Technology, Korea
Guohong Fu	Heilongjiang University, China
Choochart Haruechaiyasak	National Electronics and Computer Technology Center, Thailand
Xuanjing Huang	Fudan University, China
Kentaro Inui	Nara Institute of Science and Technology, Japan
Seung-Shik Kang	Kookmin University, Korea
Chunyu Kit	City University of Hong Kong, Hong Kong
Olivia Oi Yee Kwong	City University of Hong Kong, Hong Kong
Sobha L.	Anna University - KBC, India

Wai Lam	The Chinese University of Hong Kong, Hong Kong
Jong-Hyeok Lee	Pohang University of Science and Technology, Korea
Sujian Li	Peking University, China
Haizhou Li	Institute for Infocomm Research, Singapore
Wenjie Li	The Hong Kong Polytechnic University, Hong Kong
Ting Liu	Harbin Institute of Technology, China
Qun Liu	Chinese Academy of Sciences, China
Qing Ma	Ryukoku University, Japan
Diego Molla	Macquarie University, Australia
Masaaki Nagata	NTT Communication Science Laboratories, Japan
Manabu Okumura	Tokyo Institute of Technology, Japan
Jong Cheol Park	Korea Advanced Institute of Science and Technology, Korea
Sudeshna Sarkar	Indian Institute of Technology Kharagpur, India
Yohei Seki	Toyohashi University of Technology, Japan
Jian Su	Institute for Infocomm Research, Singapore
Zhifang Sui	Peking University, China
Le Sun	Chinese Academy of Sciences, China
Bin Sun	Peking University, China
Thanaruk Theeramunkong	Sirindhorn International Institute of Technology, Thailand
Takehito Utsuro	University of Tsukuba, Japan
Vasudev Varma	International Institute of Information Technology, India
Yunqing Xia	Tsinghua University, China
Ruifeng Xu	City University of Hong Kong
Min Zhang	Institute for Infocomm Research, Singapore
Tiejun Zhao	Harbin Institute of Technology, China
Hai Zhao	City University of Hong Kong, Hong Kong
Guodong Zhou	Suzhou University, China
Jingbo Zhu	Northeastern University, China

Hosted By

Chinese and Oriental Languages Computer Society (COLCS)

Organized By

The Hong Kong Polytechnic University, Hong Kong
The Chinese University of Hong Kong, Hong Kong
City University of Hong Kong, Hong Kong

Supported By

Asian Federation of Natural Language Processing (AFNLP)
Department of Computing, The Hong Kong Polytechnic University, Hong Kong
Department of Systems Engineering and Engineering Management,
The Chinese University of Hong Kong, Hong Kong
Centre for Language Technology, Macquarie University, Australia