Program of ROCLING XVII

地點：成功大學電機系館
議程會場：Room A：一樓繁城講堂 (92127)
Room B：一樓靄雲廳 (92119)
用餐地點：Room C：地下一樓會議室 (92X33)

September 15		
Time	Session	Chair
08:30-09:00	Registration	
09:00-09:10	Opening Ceremony (Room A) 王駿發教授/吳宗憲教授	
09:10-10:10	Keynote Speech：Prof. Jian-Yun Nie (Room A) Combining Linguistic Resources and Statistical Language Modeling for Information Retrieval	簡立峰教授
10:10-10:40	Coffee Break	
10:40-12:00	S1: Language/Speaker Identification (Room A)	余明興教授
	S2: Information Retrieval/Extraction & Summarization (Room B)	曾元顯教授
12:00-13:20	Lunch (Room C) 中華民國計算語言學學會會員大會 (Room A)	
13:20-14:00	Invited Speech – I：陳克健教授 (Room A) Lexical Knowledge Representation and Semantic Composition	吳宗憲教授
14:00-14:10	Break	
14:10-15:10	Tutorial：黃居仁教授 (Room A) Sinica BOW	盧文祥教授
	S3: Language Learning (Room B)	高照明教授
15:10-15:30	Coffee Break	
15:30-16:50	Panel Discussion (Room A) 台灣在新世代自動語音辨識技術研究應有之做法	余孝先博士
17:00-18:30	Tainan City Tour	
18:30-20:30	Banquet	

September 16		
Time	Session	Chair
09:00-09:40	Invited Speech – II：張俊盛教授 (Room A) In Search of Next Killer Apps for Natural Language and Speech Processing	簡仁宗教授
09:40-10:00	Coffee Break	
10:00-12:00	S4: Speech Analysis/Synthesis (Room A)	王逸如教授
	S5: Syntax/Semantics (Room B)	陳柏琳教授
12:00-13:00	Lunch	
13:00-14:20	S6: Speech Recognition/Enhancement (Room A)	古鴻炎教授
	S7: Multilingual/Multimedia Processing (Room B)	張景新教授
14:20-	Closing	

Keynote Speech

Combining Linguistic Resources and Statistical Language Modeling for Information Retrieval

Jian-Yun Nie
DIRO, University of Montreal, Canada
nie@iro.umontreal.ca

Abstract

In recent years, Information Retrieval (IR) has extensively explored the utilization of statistical Language Modeling (LM). This exploration has been very successful at least from two points of view: 1). From a theoretical point of view, LM offers a solid and interesting framework to IR. It allows explain several empirical weighting factors used in the past, such as TF and IDF. 2). In practice, LM has produced competitive or superior retrieval effectiveness to the state of the art systems (such as Okapi systems using vector space model). The basic idea of IR based on LM is to construct a LM for each document (a probability function); then try to estimate the likelihood that each document model can generates the query. The more a document model is likely to generate the query, the higher the document will be ranked.

We notice, however, that most of the language models used previously are unigram models. This means that words are assumed to be independent, which obviously is not true in reality. However, it is difficult and often ineffective to increase LM to bigram or trigram models. Indeed, the previous experiments showed that bigram models only outperform unigram models slightly, however, with a much higher computational cost. Then a natural alternative is to combine linguistic analysis and resources into LM, in such a way that LM can capture some of the linguistic phenomena correctly.

LM can be augmented by linguistic analysis and resources in two different ways:

- Instead of using unigrams, one can try to determine dependences between words in a sentence using a statistical or linguistic parsing, and incorporate them into LM. In this way, words are no longer considered to be independent. The resulting model is different from a bigram model in which every pair of adjacent words is blindly considered to be dependent. Instead, dependences can be created between pairs of words that are more distant.
- In classical LM, smoothing should be used to deal with the problem of data sparseness. However, the classical smoothing only redistributes a portion of the probability mass to the words that do not appear in the document, and this redistribution is done according to the distribution of words in the whole document collection. No relationship between words is considered. As a consequence, after smoothing a document about "database", the term "computer" may well have comparable probability to the term "water" (suppose that the word "computer" does not appear in that document). In fact, from some of the linguistic resources (e.g. Wordnet) or general knowledge, one can know that there are strong relationships between some words (e.g. *computer* and *database*). Then it is possible to integrate such relationships into the smoothing process in such a way that "computer" will be attributed a higher probability in the smoothing process (that we may call a semantic smoothing).

In our recent studies, we implemented the above two approaches and tested in IR experiments. It is shown that the integration of both a statistical parsing and a linguistic resource such as Wordnet is highly useful in IR.

Invited Speech I

Lexical Knowledge Representation and Semantic Composition

Keh-Jiann Chen
Institute of Information Science
Academia Sinica, Taipei, Taiwan
kchen@iis.sinica.edu.tw

Abstract

Although natural languages provide the means to denote concepts, word sense ambiguities and complexity of semantic processing make conceptual processing and natural language understanding almost impossible. To bridge the gaps, computer systems should equip with lexical knowledge bases and conceptual ontologies which are rich enough to provide the following functions.

a) Map lexical senses to concetual representations.
b) Identify synonym concepts and measure the similarity distance between two concepts.
c) Know the shared semantic features and feature differences between two concepts.
d) Provide a unique index of each concept, such that associated knowledge can be coded and accessed.
e) Utilize language independent sense encoding.
f) Make logical inferences through a conceptual property inheritance system.
g) Incorporate dynamic concept decomposition and composition mechanisms.
h) Support common sense knowledge.

In this talk, we present a universal concept representation model to achieve the above goals. It is called Extended-HowNet, which was evolved from HowNet. It extends the word sense definition mechanism of HowNet and uses WordNet synsets as vocabulary to describe concepts. Each word sense (or concept) is defined by some simpler concepts. The simple concepts used in the definitions can be further decomposed into even simpler concepts, until primitive or basic concepts are obtained. In this way, definitions can be dynamically decomposed and unified into Extended-HowNet representations at different levels. Extended-HowNet is language independent; thus, any word sense of any language can be defined and near-canonical representation can be achieved. Given any two concepts, not only their semantic distances, but also their sense similarities and differences can be derived by comparing their definitions. In addition to taxonomy links, concepts are also associated by their shared conceptual features, while fine-grain differences among near-synonyms can be differentiated by adding new features. Semantic composition became a feature unification process guided by relational and conceptual constraints.

Invited Speech II

In Search of Next Killer Apps for Natural Language and Speech Processing

Jason S. Chang
Department of Computer Science
National Tsing-Hua University
jschang@cs.nthu.edu.tw

Abstract

Language learning and tests are on everybody's mind, including students, teachers, and parents, all the time. Now, increasingly, researchers are looking into new ways to advance computer assisted language learning (CALL) and computer assisted testing (CAT) by exploiting corpora of all kinds and applying maturing technology of artificial intelligence, natural language processing, and speech processing. Reference corpora (e.g. BNC, British National Corpus and MICASE, MIchigan Corpus of Spoken English) have been exploited for teaching college level English and Academic Writing. Native and non native speakers' learner corpora have opened the door for the first time for objective study of the bottlenecks of learning a second language. Many emerging educational applications have shown to be promising and are poised to change the way we look at language learning: E-Rater, an automated program for evaluating students' free-responses has been used in such tests as GRE, while another program MOVER has been helping students analyze 'moves' of research articles in Academic Writing classes. Researchers are also making progress in creating software programs that generate test questions, for reading/writing as well as for listening/speaking, that are better than produced manually the traditional way. In short, educational applications based on NLP are becoming the next killer apps.

Table of Contents

Session 1: Language/Speaker Identification

Session 2: Information Retrieval/Extraction & Summarization

Session 3: Language Learning

Session 4: Speech Analysis/Synthesis

Session 5: Syntax/Semantics

Session 6: Speech Recognition/Enhancement

Session 7: Multilingual/Multimedia Processing