Alexander Gelbukh (Ed.)

Computational Linguistics and Intelligent Text Processing

7th International Conference, CICLing 2006
Mexico City, Mexico, February 19-25, 2006
Proceedings

Springer

Volume Editor

Alexander Gelbukh
National Polytechnic Institute (IPN)
Center for Computing Research (CIC)
Col. Zacatenco, CP 07738, D.F., Mexico
E-mail: see www.gelbukh.com

Library of Congress Control Number: Applied for

CR Subject Classification (1998): H.3, I.2.7, I.7, I.2, F.4.3

LNCS Sublibrary: SL 1 – Theoretical Computer Science and General Issues

ISSN 0302-9743
ISBN-10 3-540-32205-1 Springer Berlin Heidelberg New York
ISBN-13 978-3-540-32205-4 Springer Berlin Heidelberg New York

Springer is a part of Springer Science+Business Media

springeronline.com

Typesetting: Camera-ready by author, data conversion by Scientific Publishing Services, Chennai, India
Printed on acid-free paper SPIN: 11671299 06/3142 5 4 3 2 1 0

Preface

CICLing 2006 (www.CICLing.org) was the 7th Annual Conference on Intelligent Text Processing and Computational Linguistics. The CICLing conferences are intended to provide a wide-scope forum for discussion of the internal art and craft of natural language processing research and the best practices in its applications.

This volume contains the papers included in the main conference program (full papers) and selected papers from the poster session (short papers). Other poster session papers were included in a special issue of the journal *Research on Computing Science*; see information on this issue on the website. The previous CICLing conferences since 2001 were also published in Springer-Verlag's Lecture Notes in Computer Science (LNCS) series, vol. 2004, 2276, 2588, 2945, and 3406.

The number of submissions to CICLing 2006 was higher than that of the previous conferences: 141 full papers and 35 short papers by 480 authors from 37 countries were submitted for evaluation, see Tables 1 and 2. Each submission was reviewed by at least two independent Program Committee members. This book contains revised versions of 43 full papers (presented orally at the conference) and 16 short papers (presented as posters) by 177 authors from 24 countries selected for inclusion in the conference program. The acceptance rate was 30.4% for full papers and 45.7% for short papers.

The book is structured into two parts subdivided into 14 sections representative of the main tasks and applications of Natural Language Processing:

Computational Linguistics Research

- Lexical Resources
- Corpus-based Knowledge Acquisition
- Morphology and Part-of-Speech Tagging
- Syntax and Parsing
- Word Sense Disambiguation and Anaphora Resolution
- Semantics
- Text Generation
- Natural Language Interfaces and Speech Processing

Intelligent Text Processing Applications

- Information Retrieval
- Question Answering
- Text Summarization
- Information Extraction and Text Mining
- Text Classification
- Authoring Tools and Spelling Correction

The volume features invited papers by **Eduard Hovy** of the Information Sciences Institute, University of Southern California, **Nancy Ide** of the Vassar College, and **Rada Mihalcea** of the University of North Texas, who presented excellent keynote lectures at the conference. Publication of extended full-text invited papers in the Proceedings is a distinctive feature of CICLing conferences. What is more, in addition to presentation of their invited papers, the

Table 1. Statistics of submissions and accepted papers by country or region.

Country or region	Authors		Papers[1]	
	Subm	Accp	Subm	Accp
Algeria	2	–	1	–
Argentina	1	1	0.5	0.5
Austria	6	–	1	–
Belgium	2	1	1.2	0.2
Brazil	10	10	3.33	3.33
Canada	10	5	5	2
Chile	3	–	0.65	–
China	68	19	22	5.45
Costa Rica	1	–	0.5	–
Cuba	1	1	0.25	0.25
Czech Republic	9	2	5	1
France	12	1	5.7	1
Germany	2	1	0.53	0.33
Hong Kong	25	12	9.8	4.8
India	21	2	10	1
Ireland	4	–	1	–
Israel	3	–	1	–
Italy	3	2	1	0.5
Japan	57	18	15.25	5.5
Korea, South	67	17	29	7
Lebanon	1	–	1	–
Mexico	51	24	17.65	7.23
Netherlands	1	1	0.5	0.5
Norway	2	–	1	–
Portugal	6	6	1.5	1.5
Romania	2	–	1.5	–
Russia	5	1	2.25	0.25
Singapore	1	1	0.25	0.25
Spain	49	22	14.1	6
Sweden	2	–	2	–
Taiwan	12	3	4	1
Tunisia	3	3	1	1
Turkey	3	–	2	–
UAE	2	–	1	–
UK	3	2	0.8	0.6
USA	29	22	11.27	8
Uruguay	1	–	0.5	–
Total:	*480*	*177*	*176*	*59*

[1] Counted by authors. E.g, for a paper by 3 authors: 2 from Mexico and 1 from USA, we added $\frac{2}{3}$ to Mexico and $\frac{1}{3}$ to USA.

keynote speakers organized separate vivid informal discussions and encouraging tutorials—which is also a distinctive feature of this conference series.

The following papers received the Best Paper Awards and the Best Student Paper Award, correspondingly:

1st Place:	Shallow Case Role Annotation using Two-Stage Feature-Enhanced String Matching, by Samuel Chan;
2nd Place:	Finite State Grammar Transduction from Distributed Collected Knowledge, by Rakesh Gupta and Ken Hennacy;
3rd Place:	Automatic Acquisition of Question Reformulations for Question Answering, by Jamileh Yousefi and Leila Kosseim;
Student:	Clustering Abstracts of Scientific Texts using the Transition Point Technique, by David Pinto, Héctor Jiménez-Salazar and Paolo Rosso.

The Best Student Paper was selected out of papers with the first author being a full-time student. The authors of the awarded papers were given extended time for their presentations. In addition, the Best Presentation Award and Best Poster Award winners were selected by a ballot among the participants of the conference.

Besides its high scientific level, one of the success factors of CICLing conferences is their excellent cultural program. CICLing 2006 was held in Mexico, a wonderful country rich in culture, history, and nature. The participants of the

Table 2. Statistics of submissions and accepted papers by topic.[2]

Topic	Submitted	Accepted	
Theories and formalisms	9	2	22%
Lexical resources	29	13	44%
Statistical methods and machine learning	35	13	37%
Corpus linguistics	21	11	52%
Morphology	11	5	45%
Syntax (linguistic aspects)	12	4	33%
Parsing (technical aspects)	13	5	38%
Ambiguity resolution	16	9	56%
Word Sense Disambiguation	16	7	43%
Anaphora resolution	3	1	33%
Semantics	29	9	31%
Knowledge representation	26	4	15%
Text generation	4	4	100%
Machine translation	7	1	14%
Discourse and dialogue	8	4	50%
Natural language interfaces	10	3	30%
Speech recognition	8	3	37%
Speech synthesis	2	1	50%
Information retrieval	42	11	26%
Information extraction	25	4	16%
Text mining	26	6	23%
Summarization	6	3	50%
Text categorization	21	4	19%
Text clustering	12	5	41%
Spell checking	2	1	50%
Other: computational linguistics art and craft	12	2	16%
Other: text processing applications	38	13	34%

[2] According to the topics indicated by the authors. A paper may be assigned to more than one topic.

conference had a chance to see the solemn 2000-years-old pyramids of the legendary Teotihuacanas, a monarch butterfly wintering site where the old pines are covered with millions of butterflies as if they were leaves, a great cave with 85-meter halls and a river flowing from it, Aztec warriors dancing in the street in their colorful plumages, and the largest anthropological museum in the world; see photos at www.CICLing.org.

I want to thank all people involved in the organization of this conference. In the first place these are the authors of the papers constituting this book: it is the excellence of their research work that gives value to the book and sense to the work of all other people involved. I thank the Program Committee members for their hard and very professional work on reviewing so many submissions in a short time. Very special thanks go to Manuel Vilares and his group, John Tait and his group, Nicolas Nikolov, Rada Mihalcea, Ted Pedersen, and Oana

Postolache for their invaluable support in the reviewing process. The Best Paper Award selection working group included Alexander Gelbukh, Eduard Hovy, Rada Mihalcea, Ted Pedersen, and Yorick Wiks.

The entire submission, reviewing, and selection process, as well as putting together the Proceedings, was supported for free by the EasyChair system (www.EasyChair.org); I express my gratitude to its author Andrei Voronkov for his constant support and help. I also express my most cordial thanks to the members of the local Organizing Committee for their considerable contribution to making this conference become a reality, and to our sponsoring organization—the Center for Computing Research (CIC) of the National Polytechnic Institute (IPN), Mexico—for hosting the conference. Last but not least, I deeply appreciate the Springer-Verlag staff's patience and help in editing this volume—it is always a great pleasure to work with them.

December 2005 Alexander Gelbukh

Organization

CICLing 2006 was organized by the Natural Language and Text Processing Laboratory of the Center for Computing Research (CIC, www.cic.ipn.mx) of the National Polytechnic Institute (IPN), Mexico.

Program Chair

Alexander Gelbukh

Program Committee

Eneko Agirre
Christian Boitet
Igor Bolshakov
Nicoletta Calzolari
John Carroll
Dan Cristea
Barbara Di Eugenio
Gregory Grefenstette
Linda van Guilder
Cătălina Hallett
Yasunari Harada
Eduard Hovy
Nancy Ide
Diana Inkpen
Frederick Jelinek
Aravind Krishna Joshi
Martin Kay
Alma Kharrat
Adam Kilgarriff
Richard Kittredge
Kevin Knight
Alexander Koller
Grzegorz Kondrak
Sandra Kuebler
Ken Litkowski
Hugo Liu
Aurelio López López
Bernardo Magnini
Daniel Marcu
Carlos Martín-Vide
Igor Mel'čuk
Rada Mihalcea
Ruslan Mitkov
Masaki Murata
Vivi Nastase
Olga Nevzorova
Nicolas Nicolov
Sergei Nirenburg
Constantin Orasan
Manuel Palomar
Ted Pedersen
Viktor Pekar
Stelios Piperidis
James Pustejovsky
Fuji Ren
Fabio Rinaldi
Horacio Rodriguez
Vasile Rus
Ivan Sag
Franco Salvetti
Serge Sharoff
Grigori Sidorov
Thamar Solorio
Carlo Strapparava
Maosong Sun
John Tait
Benjamin Ka-yin T'sou
Felisa Verdejo
Karin Verspoor
Manuel Vilares Ferro
Yorick Wilks

Additional Referees

Mohamed Abdel Fattah
Mustafa Abusalah
Farooq Ahmad
Iñaki Alegria
Muath Alzghool
Bogdan Babych
Verginica Barbu Mititelu
Fco. Mario Barcala Rodríguez
Francesca Bertagna
Dimitar Blagoev
Hiram Calvo Castro
Anna Clark
Daoud Clarke
Andras Csomai
Victor Manuel Darriba Bilbao
Jeremy Ellman
Davide Fossati
Oana Frunza
Irbis Gallegos
Jorge Graña
Samer Hassan
David Hope
Scott K. Imig
Aminul Islam
Shih-Wen Ke
Stephan Kepser
Rob Koeling
Alberto Lavelli
Fennie Liang
Christian Loza
Xin Lu
Fernando Magán Muñoz
Raquel Martínez
Jaime Mendez
Dragos Stefan Munteanu
Crystal Nakatsu
Apostol Natsev
Matteo Negri
Michael Oakes
Octavian Popescu
Oana Postolache
Christoph Reichenbach
Francisco Ribadas Peña
German Rigau
Tarek Sherif
Radu Soricut
Chris Stokoe
Rajen Subba
Hristo Tanev
Martin Thomas
Jesus Vilares Ferro
Zhuli Xie

Organizing Committee

Hiram Calvo Castro
Hugo Coyote Estrada
Ignacio García Araoz
Alexander Gelbukh
Martín Haro Martínez
Oralia del Carmen Pérez Orozco
Marisol Pineda Pérez
Jorge Sosa Sánchez
Javier Tejada Cárcamo
Sulema Torres Ramos

Website and Contact

The website of CICLing conferences is www.CICLing.org. It contains information on the past CICLing events and satellite workshops, abstracts of all published papers, photos from all CICLing events, and video recordings of some keynote talks, as well as the information on the forthcoming CICLing event. Contact: cicling.org, gelbukh.com; more contact options can be found on the website.

Table of Contents

Computational Linguistics Research

Lexical Resources

Corpus-Based Knowledge Acquisition

Morphology and Part-of-Speech Tagging

Syntax and Parsing

Word Sense Disambiguation and Anaphora Resolution

Semantics

Text Generation

Natural Language Interfaces and Speech Processing

Intelligent Text Processing Applications

Information Retrieval

Question Answering

Text Summarization

Information Extraction and Text Mining

Text Classification

Authoring Tools and Spelling Correction

Integrating Semantic Frames from Multiple Sources

Namhee Kwon and Eduard Hovy

Information Sciences Institute
University of Southern California
Marina del Rey, CA 90292, USA
{nkwon,hovy}@isi.edu

Abstract. Semantic resources of predicate-argument structure have high potential to enable increased quality in language understanding. Several alternative frame collections exist, but they cover different sets of predicates and use different role sets. We integrate semantic frame information given a predicate verb using three available collections: FrameNet, PropBank, and the LCS database. For each word sense in WordNet, we automatically assign the corresponding FrameNet frame and align frame roles between FrameNet and PropBank frames and between FrameNet and LCS frames, and verify the results manually. The results are avilable as part of ISI's Omega ontology.

1 Introduction

With more accurate semantic analysis, systems should obtain higher performance in many applications such as machine translation, question answering, and summarization. Thanks to the release of annotated corpora with semantic argument structures and manually constructed lexical-semantic information such as FrameNet [1], PropBank [10], LCS database [3], and VerbNet [11], many models inducing semantic frames have been developed ([7], [6], [13], [17], [18]).

Such data collections cover different sets of predicates. Unfortunately, no collection covers all (or most) of the (English) predicates, and the roles and other definitional aspects of the collections differ. Due to these differences, most approaches to semantic analysis using these available resources (semantic role tagging) are specific to only one of these resources and their results are not comparable and usable over other resources.

We believe that we can build a broader and consistent semantic resource by integrating all semantic frame information from these disparate collections. The value of the integrated resource is apparent at many levels: first, as a theoretical device to highlight differences and generate further refinements in lexical semantic theory; second, as a practical resource that can be used by semantic analysis and other applications; third, as a testbed for an automatic aligning method between different resources that can also be applied to more general integration of lexical information. As more such annotated collections are created

Making Senses: Bootstrapping Sense-tagged Lists of Semantically-Related Words

Nancy Ide

Department of Computer Science, Vassar College, Poughkeepsie, New York USA
ide@cs.vassar.edu

Abstract. The work described in this paper was originally motivated by the need to map verbs associated with FrameNet 2.0 frames to appropriate WordNet 2.0 senses. As the work evolved, it became apparent that the developed method was applicable for a number of other tasks, including assignment of WordNet senses to word lists used in attitude and opinion analysis, and collapsing WordNet senses into coarser-grained groupings. We describe the method for mapping FrameNet lexical units to WordNet senses and demonstrate its applicability to these additional tasks. We conclude with a general discussion of the viability of using this method with automatically sense-tagged data.

1 Introduction

Lists of semantically-related words and phrases are heavily used in many automatic language processing tasks. A common use of such lists in recent work is in attitude or opinion analysis, where words indicative of a given semantic orientation—often, “positive” or negative” polarity—are detected to classify documents such as movie and product reviews as more or less favorable ([1], [2], [3]). Approaches include simple term counting [4] as well as training machine learning algorithms to classify documents. In machine learning approaches, semantically-related words and phrases are often used as a part of the feature set (e.g., [2], [3], [5]. NLP tasks such as event recognition also typically rely on lists of semantically-related verbs coupled with frames or patterns that are used to identify participants, etc. (e.g., [6] [7]).

Largely due to the recent upsurge in work on attitude and opinion analysis, numerous lists of semantically-related words have been made available within the language processing community. The lists are compiled using a variety of means, including extraction from existing resources such as lexicons, thesauri, and pre-compiled content category lists such as the General Inquirer [8]; automated extraction [2] [3]; and manual production; and often include hundreds or even thousands of words.

Whatever the source, available lists of semantically-related words do not identify the sense of the included items, despite the fact that many of the words are highly

Enriching Wordnets with New Relations and with Event and Argument Structures*

Raquel Amaro[1]**, Rui P. Chaves[1], Palmira Marrafa[1,2], and Sara Mendes[1]***

[1] CLG – Group for the Computation of Lexical and Grammatical Knowledge,
Center of Linguistics, University of Lisbon, Portugal
[2] Department of Linguistics of the Faculty of Arts, University of Lisbon, Portugal
{ramaro,rui.chaves,palmira.marrafa,sara.mendes}@clul.ul.pt

Abstract. This paper argues that wordnets, being concept-based computational lexica, should include information on event and argument structures. This general approach is relevant both for allowing computational grammars to cope with a number of different lexical semantics phenomena, as well as for enabling inference applications to obtain finer-grained results. We also propose new relations in order to adequately model non explicit information and cross-part-of-speech relations.

1 Introduction

Wordnets are electronic databases developed along with the same general lines of the so-called Princeton WordNet, an electronic database of English [1,2] containing nouns, verbs, adjectives, and adverbs. This database is structured as a network of relations between *synsets* (a set of roughly synonymous word forms). Several other wordnets have since been developed for many other languages and the number of relations adopted by the system has been enlarged (see for instance EuroWordNet [3]). In this paper we will show how wordnets can be integrated with a finer-grained lexical description framework in order to deal with various complex lexical semantics phenomena in a general and systematic way. Such an extension can be used both for deep lexical semantics analysis in computational grammars, and for a finer-grained linguistic knowledge-base in inference and question answering systems.

In Section 2 we will discuss the hyponymy/hypernymy relation. Following [4] we propose augmenting wordnet synset nodes with rich lexical-semantics descriptions which allow to explicitly capture the semantic inheritance patterns between hyponyms and hypernyms. We discuss some technical issues concerning this approach and provide a more general alternative view of semantic compatibility. Section 3 is dedicated to the verbal lexicon, focusing on argument

* The research reported in this paper has been partially supported by Instituto Camões, Fundação Calouste Gulbenkian, and Fundação para a Ciência e Tecnologia through the WordNet.PT, UniNet.PT, and COMPGRAM projects.

** Supported by FCT grant SFRH/BD/13875/2003.

*** Supported by FCT grant SFRH/BD/8524/2002.

Experiments in Cross-language Morphological Annotation Transfer

Anna Feldman, Jirka Hana, Chris Brew

Ohio State University
Department of Linguistics
Columbus, OH 43210-1298, USA

Abstract. Annotated corpora are valuable resources for NLP which are often costly to create. We introduce a method for transferring annotation from a morphologically annotated corpus of a source language to a target language. Our approach assumes only that an unannotated text corpus exists for the target language and a simple textbook which describes the basic morphological properties of that language is available. Our paper describes experiments with Polish, Czech, and Russian. However, the method is not tied in any way to these languages. In all the experiments we use the TnT tagger ([3]), a second-order Markov model. Our approach assumes that the information acquired about one language can be used for processing a related language. We have found out that even breathtakingly naive things (such as approximating the Russian transitions by Czech and/or Polish and approximating the Russian emissions by (manually/automatically derived) Czech cognates) can lead to a significant improvement of the tagger's performance.

1 Introduction

Genetically related languages posses a number of properties in common. For example, Czech and Russian are similar in many areas, including lexicon, morphology, and syntax (they have so-called free word-order). This paper explores the resemblances between Czech, Russian, and Polish, as well as exploits linguistic knowledge about these languages for automatic morpho-syntactic annotation without using parallel corpora or bilingual lexicons. Our experiments use these three languages; however, a broader goal of this work is to explore the general possibility of porting linguistic knowledge acquired in one language to another. This portability issue is especially relevant for minority languages with few resources.

Cross-language information transfer is not new; however, most of the existing work relies on parallel corpora (e.g. [7, 11, 12]) which are difficult to find, especially for lesser studied languages, including many Slavic languages. In our work, we explore a new avenue — We use a resource-rich language (e.g. Czech/Polish) to process a resource-poor genetically related language (e.g. Russian) without using a bilingual lexicon or a parallel corpus.

We tag Russian by combining information from a resource-light morphological analyzer ([5]) and information derived from Czech and Polish.

Sentence Segmentation Model to Improve Tree Annotation Tool

So-Young Park* Dongha Shin* and Ui-Sung Song**

*College of Computer Software & Media Technology, SangMyung University,
7 Hongji-dong, Jongno-gu, SEOUL, 110-743, KOREA
ssoya@smu.ac.kr, dshin@smu.ac.kr
**Dept. of Computer Science & Engineering, Korea University,
5-ka 1, Anam-dong, Seongbuk-ku, SEOUL, 136-701, KOREA
ussong@disys.korea.ac.kr

Abstract. In this paper, we propose a sentence segmentation model for a semi-automatic tree annotation tool using a parsing model. For the purpose of improving both parsing performance and parsing complexity without any modification of the parsing model, the tree annotation tool performs two-phase parsing for the intra-structure of each segment and the inter-structure of the segments after segmenting a sentence. Experimental results show that it can reduce manual effort about 28.3% by the proposed sentence segmentation model because an annotator's intervention related to cancellation and reconstruction remarkably decrease.

1 Introduction

A treebank is a corpus annotated with syntactic information. In order to reduce manual effort for building a treebank by decreasing the frequency of the human annotators' intervention, several approaches have tried to assign an unambiguous partial syntactic structure to a segment of each sentence. The approaches [1, 2] utilize the reliable heuristic rules written by the grammarians. However, it is too difficult to modify the heuristic rules, and to change the features used for constructing the heuristic rules [3]. One the other hand, the approaches [3, 4] use the rules which are automatically extracted from an already built treebank. Nevertheless, they place a limit on the manual effort reduction and the annotating efficiency improvement because the extracted rules are less credible than the heuristics.

In this paper, we propose a tree annotation tool using an automatic full parsing model for the purpose of shifting the responsibility of extracting the reliable syntactic rules to the parsing model. In order to improve both parsing performance and parsing complexity without any modification of the parsing model, it utilizes a sentence segmentation model so that it performs two-phase parsing for the intra-structure of each segment and the inter-structure of the segments after segmenting a sentence. Next, section 2 will describe the proposed sentence segmentation model for the tree annotation tool, and section 3 shows the experimental results. Finally, we conclude this paper in section 4.

Markov Cluster Shortest Path
Founded upon the Alibi-breaking Algorithm

Jaeyoung Jung, Maki Miyake, and Hiroyuki Akama

Tokyo Institute of Technology, Department of Human System Science
2-12-1 O-okayama, Meguro-ku, Tokyo, 152-8552 Japan
{catherina, mmiyake, akama}@dp.hum.titech.ac.jp

Abstract. In this paper, we propose a new variant of the breadth-first shortest path search called Markov Cluster Shortest Path (MCSP). This is applied to the associative semantic network to show us the flow of association between two very different concepts, by providing the shortest path of them. MCSP is obtained from the virtual adjacency matrix of the hard clusters taken as vertices after MCL process. Since each hard cluster grouped by concepts as a result of MCL has no overlap with others, we propose a method called Alibi-breaking algorithm, which calculates the adjacency matrix of them in a way of collecting their past overlapping information by tracing back to the on-going MCL loops. The comparison is made between MCSP and the ordinary shortest paths to know the difference in quality.

1 Introduction

In the leading network science, the graph structure and scale problem has risen as a renewed matter of concern. The same thing is true of the corpus or cognitive linguistics that allows us to see the world of language as a large-scale graph of words. If a word is associated in a certain sense to the other, it is told that they are connected with each other and all the words taken in this way as nodes (vertices) are linked together by a set of edges corresponding here with the lexical association. In this structure, the shortest path between two random words or concepts represents their distance in semantic networks. Steyvers et al. (2003) showed that large-scale word association data possess a small-world structure characterized by the combination of highly clustered neighborhoods and a short average path length. According to them, the average shortest path (SP) length between any two words was 3.03 in the Undirected Associative Network of Nelson et al, 4.26 in their Directed Associative Network, 5.43 in Roget's thesaurus and 10.61 in WordNet.

It also held true in Ishizaki Associative Concepts Dictionary of Japanese Words (in abbreviation, ACD), which offered us lexical association data for graph manipulation. Its average shortest path (SP) length was 3.442 in the 43 word pairs randomly chosen from it. Despite such low values, however, it took a relatively long time (according to our experiment mentioned below, more than 1 minute on average) by the usual searching method that automatically traces the shortest routes based on the word node connectivity in semantic networks. This kind of word-to-word distance measure not

Unsupervised Learning of Verb Argument Structures

Thiago Alexandre Salgueiro Pardo[1], Daniel Marcu[2], Maria das Graças Volpe Nunes[1]

[1] Núcleo Interinstitucional de Lingüística Computacional (NILC)
CP 668 – ICMC-USP, 13.560-970 São Carlos, SP, Brasil
http://www.nilc.icmc.usp.br

[2] Information Sciences Institute (ISI)
4676 Admiralty Way, Suite 1001, Marina del Rey, CA 90292
http://www.isi.edu

taspardo@gmail.com, marcu@isi.edu, gracan@icmc.usp.br

Abstract. We present a statistical generative model for unsupervised learning of verb argument structures. The model was used to automatically induce the argument structures for the 1,500 most frequent verbs of English. In an evaluation carried out for a representative sample of verbs, more than 90% of the induced argument structures were judged correct by human subjects. The induced structures also overlap significantly with those in PropBank, exhibiting some correct patterns of usage that are not present in this manually developed semantic resource.

1 Introduction

Inspired by the impact that the availability of Penn Treebank (Marcus et al., 1993; Marcus, 1994) had on syntactic parsing, several efforts have recently focused on the creation of semantically annotated resources. The annotation of verb arguments, their roles, and preferential linguistic behaviors represents a significant fraction of these efforts. The annotations that we are focusing on here pertain to the argument structures of a verb. In particular, we look for the words/concepts that constitute the arguments required by the verbs when these are used in real sentences.

The determination of verb argument structures has been shown to be a hard task for several reasons. Little agreement exists with respect to (a) how many canonical usages a verb has, (b) which arguments are really required by a verb and (c) in what order they may be realized in sentences. For instance, examples (1)-(3) show some patterns of usage for the verb *bought*.
(1) He had bought them gifts.
(2) He bought it 40 years ago.
(3) About 8 million home water heaters are bought each year.

Intuitively, one can induce from these examples that the object/thing that is bought ("gifts" in sentence (1), "it" in sentence (2), and "about 8 million home water heaters" in sentence (3)) is more likely to be a required argument for the verb than the time when the buying event occurred, since the thing bought is specified in all the cases

A methodology for extracting ontological knowledge from Spanish documents

Rafael Valencia García[1], Dagoberto Castellanos Nieves[2],
Jesualdo Tomás Fernández Breis[1], Pedro José Vivancos Vicente[1]

[1] Departamento de Informática y Sistemas. Facultad de Informática.
Universidad de Murcia 30071 Espinardo (Murcia). España
Tel: +34 968364613, Fax: +34 968364151
{valencia, jfernand, pedroviv}@um.es
[2] Facultad de Informática y Matemática,
Universidad de Holguín, Holguín, Cuba
dcn.ncd@gmail.com

Abstract. This paper presents a semi-automatic approach for extracting knowledge from natural language texts in Spanish. The knowledge is acquired and learned through the combination of NLP techniques for analyzing text fragments, the ontological technology for representing knowledge and MCRDR, a case-based reasoning methodology. This approach has been applied in the oncology domain and the results of this application are discussed in this work.

1 INTRODUCTION

Spanish is the official language of a significant amount of countries, and it has millions of speakers world-wide. Hence, there is a huge amount of information and knowledge in Spanish documents. So, extracting knowledge from such texts would be beneficial and of great help for the Spanish speaking community. The recognition of natural language has been traditionally viewed as a linguistic issue and based on grammars. However, grammars have different drawbacks, such as the fact that they are unable of managing ambiguity, imprecision, variability, etc. In order to overcome the drawbacks of grammar approaches, we have developed a methodology for acquiring knowledge from texts in an incremental way based on knowledge engineering and natural language processing techniques. In this paper, we describe such methodology and how it is capable of extracting knowledge from pieces of Spanish free texts. The combination of knowledge engineering technologies with natural language processing techniques provides us the goodnesses of both areas. As far as knowledge engineering technologies are concerned, two have been included in the methodology, namely, ontologies and MCRDR. Let us introduce now both technologies and the reason why they are used in the methodology.

Automatically Determining Allowable Combinations of a Class of Flexible Multiword Expressions

Afsaneh Fazly, Ryan North, and Suzanne Stevenson

Department of Computer Science
University of Toronto
Toronto, ON M5S 3H5
Canada
{afsaneh,ryan,suzanne}@cs.toronto.edu

Abstract. We develop statistical measures for assessing the acceptability of a frequent class of multiword expressions. We also use the measures to estimate the degree of productivity of the expressions over semantically related nouns. We show that a linguistically-inspired measure outperforms a standard measure of collocation in its match with human judgments. The measure uses simple extraction techniques over non-marked-up web data.

1 Light Verb Constructions

Recent work in NLP has recognized the challenges posed by the rich variety of multiword expressions (MWEs) (e.g., Sag et al., 2002). One unsolved problem posed by MWEs is how they should be encoded in a computational lexicon. Many MWEs are syntactically flexible; for these it is inappropriate to treat the full expression as a single word. However, fully compositional techniques can lead to overgeneralization, because flexible MWEs are often *semi*-productive: new expressions can only be formed from limited combinations of semantically and syntactically similar component words. In order to achieve accurate lexical acquisition methods, we must determine computational mechanisms for capturing the allowable combinations of such MWEs.

Our focus here is on light verb constructions (LVCs); these are largely compositional and semi-productive MWEs having a high frequency of occurrence across many diverse languages (Karimi, 1997; Miyamoto, 2000; Butt, 2003). LVCs combine a member of a restricted set of light verbs, such as *give*, *take*, and *make* among others in English, with a wide range of complements of varying syntactic categories. We consider a common class of LVCs, in which the complement is a noun generally used with an indefinite article, as in (a–c) below:

a. Priya *took a walk* along the beach.
b. Allene *gave a smile* when she saw us.
c. Randy *made a joke* to his friends.

d. Priya *walked* along the beach.
e. Allene *smiled* when she saw us.
f. Randy *joked* to his friends.

Moreover, the complement nouns in these expressions, such as *walk*, *smile*, and *joke* in (a–c), have a stem form identical to a verb. Because the light verb is "semantically bleached" to some degree (Butt, 2003), most of the meaning of these LVCs comes from

Web-Based Measurements of Intra-Collocational Cohesion in Oxford Collocations Dictionary*

Igor A. Bolshakov[1] and Sofia N. Galicia-Haro[2]

[1] Center for Computing Research (CIC),
National Polytechnic Institute (IPN), Mexico City, Mexico
igor@cic.ipn.mx

[2] Faculty of Sciences
National Autonomous University of Mexico (UNAM)
Mexico City, Mexico
sngh@fciencias.unam.mx

Abstract. Cohesion between components of collocations is already acknowledged measurable by means of the Web, and cohesion measurements are used for some applications and extraction of new collocations. Taking a specific cohesion criterion *SCI*, we performed massive evaluations of collocate cohesion in Oxford Collocations Dictionary. For three groups of modificative collocations (adjective—noun, adverb—adjective, and adverb— verb) *SCI* distributions proved to be one-peaked and compact, with rather close mean values and standard deviations. Thus we suggest a reliable numeric criterion for extraction of collocations from the Web.

Introduction

Let us transitorily define collocations as syntactically linked and semantically compatible pairs of content words. They are rather specific for each language, so electronic collocation databases compiled beforehand are needed for many applications (text editing, foreign language learning, syntactic analysis, word sense disambiguation, detection & correction of errors etc.). Though the tools for automatic collocation extraction are being developed more than 15 years [7], large electronic collocation databases do not exist to the date for well-known languages.

The Web is acknowledged now as a huge corpus for automatic collocation extraction and this it is supposed possible with a numeric criterion of coherence between collocates [2, 6]. An application of corpus-oriented criteria to Web statistics theoretically is not grounded, since the Web counts occurrences and co-occurrences in pages, not words. Since a theory allowing to recalculate the numbers of relevant pages to the numbers of words occurred in them does not exist nowadays, we are free to use both formulas recommended for corpuses, re-conceptualizing page numbers as word numbers, and any analogous formulas operating by numbers of relevant pages.

* Work done under partial support of Mexican Government (CONACyT, SNI) and CGEPI-IPN, Mexico.

Probabilistic Neural Network Based English-Arabic Sentence Alignment

Mohamed Abdel Fattah[1], Fuji Ren[1], Shingo Kuroiwa[1]
[1] Faculty of Engineering, University of Tokushima
2-1 Minamijosanjima
Tokushima, Japan 770-8506
{mohafi, ren, kuroiwa} @is.tokushima-u.ac.jp

Abstract. In this paper, we present a new approach to align sentences in bilingual parallel corpora based on a probabilistic neural network (P-NNT) classifier. A feature parameter vector is extracted from the text pair under consideration. This vector contains text features such as length, punctuation score, and cognate score values. A set of manually aligned training data was used to train the probabilistic neural network. Another set of data was used for testing. Using the probabilistic neural network approach, an error reduction of 27% was achieved over the length based approach when applied on English-Arabic parallel documents.

1 Introduction

Recently, much work has been reported on sentence alignment using different techniques [1]. Length-based approaches (length as a function of sentence characters [2] or sentence words [3]) were the most interesting. These approaches work quite well with clean input, such as the Canadian Hansards corpus, however they do not work well with noisy document pairs. Moreover, these approaches require that the paragraph boundaries be clearly marked, which is not the case for most of document pairs. Cognate approaches have also been proposed and have been combined with length-based approaches to improve alignment accuracy [4, 5]. They have used sentence cognates such as digits, alphanumerical symbols, punctuation, and alphabetical words. However both of Simard and Thomas did not take the text length between two successive cognates (Simard case) or punctuations (Thomas case) into account which increased the system confusion that leads to execution time increase and accuracy decrease (we have avoided this drawback in this work).

In this paper we present a non-traditional approach for the sentence alignment problem. In the sentence alignment problem, we may have 1-0 (One English sentence does not match any of the Arabic sentences), 0-1, 1-1, 1-2, 2-1, 2-2, 1-3 and 3-1. There may be more categories in bi-texts, however they are rare, hence we consider only the previous mentioned categories. As illustrated above, we have eight sentence alignment categories. Hence, we can consider sentence alignment as a classification problem. This classification problem may be solved by using a probabilistic neural network classifier.

Towards the Automatic Lemmatization of 16th Century Mexican Spanish: a Stemming Scheme for the CHEM

Alfonso MEDINA-Urrea

GIL, Instituto de Ingeniería, UNAM
Ciudad Universitaria, 04510 Coyoacán, DF, MEXICO
amedinau@ii.unam.mx

Abstract. Two of the problems that should arise when developing a stemming scheme for diachronic corpora are: (1) morphological systems of natural languages may vary throughout time, and these changes are normally not documented sufficiently; and (2) they exhibit very diverse orthographic characteristics. In this short paper, a stemming strategy for a diachronic corpus of Mexican Spanish is briefly described, which partially faces up to these problems. Success rates of the method are contrasted to those of a Porter stemmer.

1 Introduction

Diachronic corpora for the Spanish language have become available for various kinds of research. Two widely known corpora are the RAE's *Corpus diacrónico del español*, CORDE (`http://www.rae.es/`), and Mark Davies' *Corpus del español* (`http://www.corpusdelespanol.org/`). Recently, a first version of the *Corpus histórico del español de México*, CHEM (`http://www.iling.unam.mx/chem/`), became available to the public for the study of the Spanish used in Mexico from the arrival of Europeans to the 19th century.

Many tools for the exploitation and analysis of corpora require a lemmatization process, which is often reduced to simple stemming or graphical word truncation to eliminate inflections. Simple techniques such as the Porter algorithm [1] are regularly applied to corpora of many languages, but they require knowledge of their morphology. Fortunately, in comparison with other languages, Spanish morphology has changed relatively little during the last five centuries. So, a Porter stemmer for today's Spanish could presumably be applied to those centuries in order to accomplish inflection removal. However, given that techniques exist which can be used for stemming without having to code morphological knowledge into the algorithm, it is worthwhile to compare them to the Porter method in order to appreciate what scheme would be better for the CHEM.

In this short paper, the stemming strategy devised for this corpus is described and contrasted with an implementation of the Porter stemmer.[1] The strategy

[1] Various implementations of the Porter algorithm for Spanish are available (based on `http://snowball.tartarus.org/`). In this experiment a version for contemporary Spanish developed at GIL-IINGEN-UNAM, was used.

Word Frequency Approximation for Chinese without Using Manually Annotated Corpus

Sun Maosong[1] , Zhang Zhengcao[1] , Benjamin K Y T'sou[2], Lu Huaming[3]

[1] The State Key Laboratory of Intelligent Technology and Systems,
Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China
sms@tsinghua.edu.cn
[2] Language Information Sciences Research Center, City University of Hong Kong
rlbtsou@cityu.edu.hk
[3] School of Business, Beijing Institute of Machinery, Beijing 100085, China
huaminglu@sohu.com

Abstract. Word frequencies play important roles in a variety of NLP-related applications. Word frequency estimation for Chinese is a big challenge due to characteristics of Chinese, in particular word-formation and word segmentation. This paper concerns the issue of word frequency estimation in the condition that we only have a Chinese wordlist and a raw Chinese corpus with arbitrarily large size, and do not perform any manual annotation to the corpus. Several realistic schemes for approximating word frequencies under the framework of STR (frequency of string of characters as an approximation of word frequency) and MM (Maximal matching) are presented. Large-scale experiments indicate that the proposed scheme, MinMaxMM, can significantly benefit the estimation of word frequencies, though its performance is still not very satisfactory in some cases.

1 Introduction

Word frequencies play important roles in a variety of NLP-related applications, for example, TF in information retrieval. The estimation of word frequencies for English is very easy – it can be done by running a simple program to count word occurrences in a (in fact, any arbitrarily huge) corpus. In case of Chinese where no explicit word boundaries like spaces exist between words in texts, the task becomes very complex.

In general, a fully correct word-segmented Chinese corpus is a prerequisite for calculating word frequencies (Liu 1973). However, we face two difficulties in this respect. The first one is such a 'fully correct' corpus, or, a corpus with ideal segmentation consistency, is extremely difficult to obtain due to a main characteristic of Chinese word-formation: the borders between morphemes, words, and phrases of Chinese are fuzzy in nature (Dai 1992; Chen 1994), though the definition for 'word' from the linguistic perspective seems very clear (Zhu 1982;Tang 1992). A large number of linguistic constituents could be regarded as words by some linguists whereas be regarded as phrases by others (even for a specific linguist, his feeling to some constituents may change in between from time to time), resulting in serious

Abbreviation Recognition with MaxEnt Model

Chunyu Kit, Xiaoyue Liu, and Jonathan J. Webster

Department of Chinese, Translation and Linguistics
City University of Hong Kong, 83 Tat Chee Ave., Kowloon, Hong Kong
{ctckit,xyliu0,ctjjw}@cityu.edu.hk

Abstract. Abbreviated words carry critical information in the literature of many special domains. This paper reports our research in recognizing dotted abbreviations with MaxEnt model. The key points in our work include: (1) allowing the model to optimize with as many features as possible to capture the text characteristics of context words, and (2) utilizing simple lexical information such as sentence-initial words and candidate word length for performance enhancement. Experimental results show that this approach achieves impressive performance on the WSJ corpus.

1 Introduction

The literature in many special domains, e.g., biomedical, has been growing rapidly in recent years with a large number of abbreviations carrying critical information, e.g., proper names and terminology. There is an increasing interest in practical techniques for identifying abbreviations from plain texts.

There are several typical forms of abbreviation, including acronyms, blending, and dotted strings. Previous research [2, 7] illustrated significant success in identifying and pairing short form terms, referred to as abbreviations, most of which are acronyms, and their original long forms, referred to as definitions, e.g., `<HIV, Human Immunodeficiency Virus>`. This paper is intended to report our recent work to apply the Maximum Entropy (MaxEnt) model to identify abbreviations in another form, i.e., dotted strings, e.g., "abbr." for "abbreviation", "Jan." for "January". Popular abbreviations of this kind such as "Mr.", "Dr.", "Prof.", "Corp." and "Ltd." are available from an ordinary dictionary. There is no point to invent any complicated techniques for recognizing them. The availability of such a sample set, however, gives us great convenience to evaluate the performance of a learning model on recognizing abbreviations with a particular surface form. The significance of this approach lies in the plausibility that similar methodology can be applied to abbreviations with some other common surface characteristics, e.g., in parentheses.

Aiming at this objective, we intend to allow the MaxEnt model to optimize with as many features as possible to capture the text form characteristics of context words and with some special features to utilize simple lexical information such as candidate word length and sentence-initial words that can be derived from the training data straightforwardly. Section 2 below presents feature selection for MaxEnt model training, and Sect. 3 the experiments for evaluation. Section 4 concludes the paper in terms of experimental results.

An Efficient Multi-Agent System Combining POS-Taggers for Arabic Texts

Chiraz Ben Othmane Zribi[1], Aroua Torjmen[1], Mohamed Ben Ahmed[1]

[1] RIADI laboratory, National School of Computer Sciences, 2010,
University of La Manouba, Tunisia
{Chiraz.benothmane, Aroua.torjmen,
Mohamed.benahmed}@riadi.rnu.tn

Abstract. In this paper, we address the problem of Part-Of-Speech tagging of Arabic texts with vowel marks. After the description of the specificities of Arabic language and the induced difficulties on the task of POS-tagging, we propose an approach combining several methods. One of these methods, based on sentences patterns, is original and very attractive. We present, afterward, the multi-agent architecture that we adopted for the conception and the realization of our POS-tagging system. The multi-agent architecture is justified by the need for collaboration, parallelism and competition between the different agents. Finally, we expose the implementation and the evaluation of the system implemented.

1 Introduction

The process of Part-Of-Speech tagging was widely automated for English and French and for many others European languages giving a rate of accuracy ranging from 95 % to 98 %. We find on the Web, many tagged corpora as well as programs of POS-tagging for these languages. The methods used by these POS-taggers are various, namely stochastic approaches such as the Hidden Markov Model [1], the decision trees [2], the maximum entropy model [3], rules-based approaches inspired in their majority of the transformation rules-based POS-tagging [4], hybrid approaches [5] (statistics and rules-based), or combined ones [6] and [7].

Unfortunately, the situation is different for Arabic as there are neither POS-taggers nor tagged corpora available. Nevertheless, some Arabic POS-taggers [8], [9] and [10] started to appear with an accuracy going from 85% to 90% on average for texts with vowel marks and by about 65% for texts without vowel marks.

This gap noted for Arabic language is especially due to, its particular characteristics, which, involve firstly, a rate of grammatical ambiguity relatively more significant than for other languages, and secondly, make impossible the application of existing POS-taggers without any change. Thus, obtaining improving accuracy remains a challenge to reach for Arabic language.

Accordingly, we propose a POS-tagging system for Arabic texts. Due to the complexity of the problem, and in order to decrease grammatical ambiguity, we have restricted the scope of our investigation: we only treat texts with vowels marks.

A Comparative Evaluation of a New Unsupervised Sentence Boundary Detection Approach on Documents in English and Portuguese

Jan Strunk[1], Carlos N. Silla Jr.[2], and Celso A. A. Kaestner[2]

[1] Sprachwissenschaftliches Institut, Ruhr-Universität Bochum,
44780 Bochum, Germany
strunk@linguistics.rub.de
[2] Pontifical Catholic University of Paraná,
Rua Imaculada Conceição 1155, 80215-901 Curitiba, Brazil
{silla,kaestner}@ppgia.pucpr.br

Abstract. In this paper, we describe a new unsupervised sentence boundary detection system and present a comparative study evaluating its performance against different systems found in the literature that have been used to perform the task of automatic text segmentation into sentences for English and Portuguese documents. The results achieved by this new approach were as good as those of the previous systems, especially considering that the method does not require any additional training resources.

1 Introduction

We are living today in an era of information overload. The web alone contains about 170 terabytes of information, which is roughly 17 times the size of the printed material in the Library of Congress of the USA; cf. [1]. However, it is becoming more and more difficult to use the available information. Many problems such as the retrieval and extraction of information and the automatic summarization of texts have become important research topics in computer science. The use of automatic tools for the treatment of information has become essential to the user because without those tools it is virtually impossible to exploit all the relevant information available on the Web.

One pre-processing component that is essential to most text-based systems is the automatic segmentation of a text into sentences. Existing systems for sentence boundary detection mostly either use a set of heuristics or a supervised machine learning approach. The drawback of both these approaches is that adapting them to new languages can be time and resource intensive. In the first case, it is necessary to adapt the rules to the new language. In the second case, a new training corpus has to be tagged manually for retraining.

In this paper, we compare a new unsupervised approach to sentence boundary detection by Kiss & Strunk [2] with the results of a previous evaluation of three

A General and Multi-lingual Phrase Chunking Model Based on Masking Method

Yu-Chieh Wu[1], Chia-Hui Chang[1], and Yue-Shi Lee[2]

[1] Department of Computer Science and Information Engineering, National Central University,
No.300, Jhong-Da Rd., Jhongli City, Taoyuan County 32001, Taiwan, R.O.C.
bcbb@db.csie.ncu.edu.tw, chia@csie.ncu.edu.tw
[2] Department of Computer Science and Information Engineering, Ming Chuan University,
No.5, De-Ming Rd, Gweishan District, Taoyuan 333, Taiwan, R.O.C.
leeys@mcu.edu.tw

Abstract. Several phrase chunkers have been proposed over the past few years. Some state-of-the-art chunkers achieved better performance via integrating external resources, e.g., parsers and additional training data, or combining multiple learners. However, in many languages and domains, such external materials are not easily available and the combination of multiple learners will increase the cost of training and testing. In this paper, we propose a mask method to improve the chunking accuracy. The experimental results show that our chunker achieves better performance in comparison with other deep parsers and chunkers. For CoNLL-2000 data set, our system achieves 94.12 in F rate. For the base-chunking task, our system reaches 92.95 in F rate. When porting to Chinese, the performance of the base-chunking task is 92.36 in F rate. Also, our chunker is quite efficient. The complete chunking time of a 50K words document is about 50 seconds.

1 Introduction

Automatic text chunking aims to determine non-overlap phrases structures (chunks) in a given sentence. These phrases are non-recursive, i.e., they cannot be included in other chunks [1]. Generally speaking, there are two phrase chunking tasks, including text chunking (shallow parsing) [15], and noun phrase (NP) chuncking [16]. The former aims to find the chunks that perform partial analysis of the syntactic structures in texts [15], while the later aims to identify the initial portions of non-recursive noun phrase, i.e., the first level noun phrase structures of the parsing trees [17] [19]. In this paper, we extend the NP chunking task to arbitrary phrase chunking, i.e., base-chunking. In comparison, shallow parsing extracts not only the first level but also the other level phrase structures of the parsing tree into the flat non-overlap chunks.

Chunk information of a sentence is usually used to present syntactic relations in texts. In many Natural Language Processing (NLP) areas, e.g., chunking-based full parsing [1] [17] [24], clause identification [3] [19], semantic role labeling (SRL) [4], text categorization [15] and machine translation, the phrase structures provide down-stream syntactic features for further analysis. In many cases, an efficient and

UCSG Shallow Parser

G. Bharadwaja Kumar, Kavi Narayana Murthy

Department of Computer and Information Sciences
University of Hyderabad, India
email: knmuh@yahoo.com,g_vijayabharadwaj@yahoo.com

Abstract. Recently, there is an increasing interest in integrating rule based methods with statistical techniques for developing robust, wide coverage, high performance parsing systems. In this paper[1], we describe an architecture, called UCSG shallow parser architecture, which combines linguistic constraints expressed in the form of finite state grammars with statistical rating using HMMs built from a POS-tagged corpus and an A* search for global optimization for determining the best shallow parse for a given sentence. The primary aim of the design of the UCSG parsing architecture is developing a judicious combination of linguistic and statistical methods to develop wide coverage robust shallow parsing systems, without the need for large scale manually parsed training corpora. The UCSG architecture uses a grammar to specify all valid structures and a statistical component to rate and rank the possible alternatives, so as to produce the best parse first without compromising on the ability to produce all possible parses. The architecture supports bootstrapping with an aim to reduce the need for parsed training corpora. The complete system has been implemented in Perl under Linux. In this paper we first describe the UCSG shallow parsing architecture and then focus on the evaluation of the UCSG finite state grammar for the chunking task for English. Recall of 91.16% and 93.73% have been obtained on the Susanne parsed corpus and CoNLL 2000 chunking task test data set respectively. Extensive experimentation is under way to evaluate the other modules.

Key Words:- Chunking, Shallow Parsing, Finite State Grammar, HMM, A* search, UCSG Architecture

1 Introduction

Although a lot of work has gone into developing full syntactic parsers, high performance, wide coverage syntactic parsing has remained a difficult challenge [1]. In recent times, there has been an increasing interest in wide coverage and robust but partial or shallow parsing systems. Shallow parsing is the task of recovering only a limited amount of syntactic information from natural language sentences. Often shallow parsing is restricted to finding phrases in sentences, in which case it is also called chunking. Steve Abney[2], has described chunking as

[1] The research work reported here was supported in part by the University Grants Commission under the UPE scheme

Evaluating the Performance of the Survey Parser with the NIST Scheme

Alex Chengyu Fang

Department of Chinese, Translation and Linguistics
City University of Hong Kong, Tat Chee Avenue, Kowloon, Hong Kong
acfang@cityu.edu.hk

Abstract. Different metrics have been proposed for the estimation of how good a parser-produced syntactic tree is when judged by a correct tree from the treebank. The emphasis of measurement has been on the number of correct constituents in terms of constituent labels and bracketing accuracy. This article proposes the use of the NIST scheme as a better alternative for the evaluation of parser output in terms of *correct match*, *substitution*, *deletion*, and *insertion*. It describes an experiment to measure the performance of the Survey Parser that was used to complete the syntactic annotation of the International Corpus of English. This article will finally report empirical scores for the performance of the parser and outline some future research.

1 Introduction

Different metrics have been proposed and all aim at the estimation of how good a parse tree is when judged by a correct tree from the Treebank (see [1], [2], [3], [4], and [5]). The emphasis of measurement has been on the number of correct constituents either in terms of constituent labels, such as *labelled match*, *precision*, and *recall*, or in terms of bracketing such as *bracketed match*. Together with *crossing brackets*, these measures indicate the number of correct and wrong matches in the parse tree. However, these measures outlined above do not constitute a satisfactory assessment. We may well imagine a parse tree with only two correct constituents scoring a high rate in terms of labelled and bracketed matches, crossing brackets, precision, and recall while deletions and insertions of nodes and associated labels could render the parse tree totally different from the correct one.

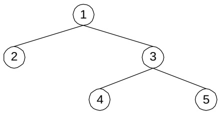

Fig. 1. *A correct tree*

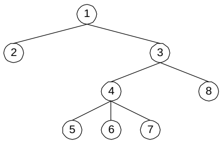

Fig. 2. *A parser-produced tree*

Sequences of Part of Speech Tags vs. Sequences of Phrase Labels

How Do They Help in Parsing?

Gabriel Infante-Lopez[1] and Maarten de Rijke[2]

[1] FaMAF, Universidad Nacional de Córdoba, Córdoba, Argentina
gabriel@famaf.unc.edu.ar
[2] Informatics Institute, University of Amsterdam, The Netherlands
mdr@science.uva.nl

Abstract. We compare the contributions made by sequences of part of speech tags and sequences of phrase labels for the task of grammatical relation finding. Both are used for grammar induction, and we show that English labels of grammatical relations follow a very strict sequential order, but not as strict as POS tags, resulting in better performance of the latter on the relation finding task.

1 Introduction

Some approaches to parsing can be viewed as a simple context free parser with the special feature that the context free rules of the grammar used by the parser do not exist a priori [**?**,**?**,**?**]. Instead, there is a device for generating bodies of context free rules on demand. Collins [**?**] and Eisner [**?**] use Markov chains as the generative device, while Infante-Lopez and De Rijke [**?**] use the more general class of probabilistic automata. These devices are induced from sample instances obtained from tree-banks. The learning strategy consists of coping all bodies of rules inside the Penn Tree-bank (PTB) to a bodies of rules sample bag which is then treated as the sample bag of an *unknown* regular language. This unknown regular language is to be induced from the sample bag, which is, later on, used for generating new bodies of rules.

Usually, the induced regular language is described by means of a probabilistic automata. The quality of the resulting automata depends on many things; the alphabet of the target regular language being one. At least two such alphabets have been considered in the literature: Part of Speech (POS) tags and grammatical relations (GRs), where the latter are labels describing the relation between the main verb and its dependents; they can be viewed as a kind of non-terminal labels. Using one or the other alphabets for grammar induction might produce different results on the overall parsing task. Which of the two produces "better" automata, that produce "better rules," which in turn lead to "better" parsing scores? This is our main research question in this paper.

Let us provide some further motivation and explanations. In order to obtain phrase structures like the ones retrieved in [**?**], the dependents of a POS tag should consist

Verb Sense Disambiguation using Support Vector Machines: impact of WordNet-extracted Features

Davide Buscaldi, Paolo Rosso, Ferran Pla,
Encarna Segarra, and Emilio Sanchis Arnal

Dpto. Sistemas Informáticos y Computación,
Universidad Politécnica de Valencia,
Valencia, Spain
{dbuscaldi, prosso, fpla, esegarra, esanchis}@dsic.upv.es

Abstract. The disambiguation of verbs is usually considered to be more difficult with respect to other part-of-speech categories. This is due both to the high polysemy of verbs compared with the other categories, and to the lack of lexical resources providing relations between verbs and nouns. One of such resources is WordNet, which provides plenty of information and relationships for nouns, whereas it is less comprehensive with respect to verbs. In this paper we focus on the disambiguation of verbs by means of Support Vector Machines and the use of WordNet-extracted features, based on the hyperonyms of context nouns.

1 Introduction

Word Sense Disambiguation (WSD) is an open problem in the field of Natural Language Processing (NLP). The resolution of lexical ambiguity that appears when a given word in a context has several different meanings is commonly referred as Word Sense Disambiguation. Supervised approaches to WSD usually perform better than unsupervised ones [4]. Results of the recent Senseval-3[1] contest attest this supremacy; moreover, recent results of the application of Support Vector Machines (SVM), a well-known supervised learning technique, to the Word Sense Dismbiguation task seem promising [3].

Some interesting results have been obtained recently in the supervised disambiguation of verbs [1], by using context-extracted features and a multi-class learning architecture. The disambiguation method described in this paper replicates the feature extraction model proposed in [1], with the addition of WordNet [5] extracted features, while using a SVM-based learning architecture. The system was tested over a subset of the Senseval-3 Lexical Sample corpus.

2 Support Vector Machines

The SVM [6] performs optimization to find a hyperplane with the largest margin that separates training examples into two classes. A test example is classified

[1] http://www.senseval.org

Preposition Senses: Generalized Disambiguation Model

Chutima Boonthum, Shunichi Toida, and Irwin Levinstein

Department of Computer Science, Old Dominion University
Norfolk, Virginia 23529 USA
{cboont, toida, ibl}@cs.odu.edu

Abstract. Our previous study on disambiguating the preposition "with" (using WordNet for hypernym and meronym relations, LCS for verb and preposition lexical information, and features of head and complement) looked promising enough to warrant study for other prepositions. Through investigation of ten frequently used prepositions, this paper describes general senses of prepositions and sense-case definitions, introduces a novel generalized sense disambiguation model, and demonstrates how this benefits a paraphrase recognition system.

1 Introduction

Why is preposition sense disambiguation important in a paraphrase recognition system? When two expressions describe the same situation, each is considered to be a paraphrase of the other. Various authorities have mentioned the following paraphrase patterns: using synonyms, changing part-of-speech, reordering ideas, breaking a sentence into smaller ones, substituting a word with its definition, and using different sentence structures. Prepositions play a significant role in changing sentence structures more than other paraphrase patterns. Consider the following sentences:

(a) "John builds a house *with* a hammer."
(b) "John *uses* a hammer *to* build a house."
(c) "John builds a house *by using* a hammer."
(d) "A house is built *by* John who *uses* a hammer."
(e) "A house is built *by* John *using* a hammer."

Although these sentences convey the same meaning, they have different syntactic structures and use different prepositions. Sentence (a) uses 'with' to indicate an instrument used to complete an action while (b), (c), (d), and (e) have the verb 'use' to indicate a use of an instrument. Sentences (d) and (e) are in the passive voice and they use the preposition 'by' to indicate an agent (who performs the action.) Sentence (c) uses 'by' to indicate a secondary action of this agent in completing the primary action. 'By' can be omitted in (c) and the sentence still has the same meaning.

(f) "John builds a house *with* a kitchen."
(g) "John builds a house *that has* a kitchen."
(h) "John builds a house *having* a kitchen."
(i) "A house is built *by* John *with* a kitchen."

An Unsupervised Language Independent Method of Name Discrimination Using Second Order Co-Occurrence Features

Ted Pedersen[1], Anagha Kulkarni[1], Roxana Angheluta[2],
Zornitsa Kozareva[3], and Thamar Solorio[4]

[1] University of Minnesota, Duluth, USA
[2] Katholieke Universiteit Leuven, Belgium
[3] University of Alicante, Spain
[4] University of Texas at El Paso, USA

Abstract. Previous work by Pedersen, Purandare and Kulkarni (2005) has resulted in an unsupervised method of name discrimination that represents the context in which an ambiguous name occurs using second order co–occurrence features. These contexts are then clustered in order to identify which are associated with different underlying named entities. It also extracts descriptive and discriminating bigrams from each of the discovered clusters in order to serve as identifying labels. These methods have been shown to perform well with English text, although we believe them to be language independent since they rely on lexical features and use no syntactic features or external knowledge sources. In this paper we apply this methodology in exactly the same way to Bulgarian, English, Romanian, and Spanish corpora. We find that it attains discrimination accuracy that is consistently well above that of a majority classifier, thus providing support for the hypothesis that the method is language independent.

1 Introduction

Purandare and Pedersen (e.g., [9], [10]) previously developed an unsupervised method of word sense discrimination that has also been applied to name discrimination by Pedersen, Purandare, and Kulkarni [8]. This method is characterized by a reliance on lexical features, and avoids the use of syntactic or other language dependent information. This is by design, since the method is intended to port easily and effectively to a range of languages. However, all previous results with this method have been reported for English only.

In this paper, we evaluate the hypothesis that this method of name discrimination is language independent by applying it to name discrimination problems in Bulgarian, Romanian, and Spanish, as well as in English.

Ambiguity in names of people, places and organizations is an increasingly common problem as online sources of information grow in size and coverage. For example, Web searches for names frequently locate different entities that share

Extracting Key Phrases to Disambiguate Personal Names on the Web

Danushka Bollegala[1], Yutaka Matsuo[2], and Mitsuru Ishizuka[1]

[1] University of Tokyo
{danushka,ishizuka}@miv.t.u-tokyo.ac.jp
[2] AIST
y.matsuo@carc.aist.go.jp

Abstract. When you search for information regarding a particular person on the web, a search engine returns many pages. Some of these pages may be for people with the same name. How can we disambiguate these different people with the same name? This paper presents an unsupervised algorithm which produces key phrases for the different people with the same name. These key phrases could be used to further narrow down the search, leading to more person specific unambiguous information. The algorithm we propose does not require any biographical or social information regarding the person. Although there are some previous work in personal name disambiguation on the web, to our knowledge, this is the first attempt to extract key phrases to disambiguate the different persons with the same name. To evaluate our algorithm, we collected and hand labeled a dataset of over 1000 Web pages retrieved from Google using personal name queries. Our experimental results shows an improvement over the existing methods for namesake disambiguation.

1 Introduction

The Internet has grown into a collection of billions of web pages. One of the most important interfaces to this vast information are web search engines. We send simple text queries to search engines and retrieve web pages. However, due to the ambiguities in the queries and the documents, search engines return lots of irrelevant pages. In the case of personal names, we may receive web pages to other people with the same name (*namesakes*). However,the the different namesakes appear in quite different contexts. For example if we search for *Michael Jackson* in Google, among the top hundred hits we get a beer expert and a gun dealer along with the famous singer. However, the context in which the singer appears is quite different from his namesakes. However, context associated with a personal name is difficult to identify. In cases where the entire web page is about the person under consideration, the context could be the complete page. On the other hand the context could be few sentences having the specified name. In this paper we explore a method which uses terms extracted from web pages to represent the context of namesakes. For example, in the case of Michael Jackson, terms such as *music, album, trial* associate with the famous singer, whereas we

Chinese Noun Phrase Metaphor Recognition with Maximum Entropy Approach[1]

Zhimin Wang, Houfeng Wang, Huiming Duan, Shuang Han, and Shiwen Yu

Department of Computer Science and Technology
Institute of Computational Linguistics, Peking University, 100871, Beijing, China
{wangzm, wanghf, duenhm, yusw }@pku.edu.cn

Abstract. This paper presents a maximum entropy (ME)-based model for Chinese noun phrase metaphor recognition. The metaphor recognizing process will be viewed as a classification task between metaphor and literal meaning. Our experiments show that the metaphor recognizer based on the ME method is significantly better than the Example-based methods within the same context windows. In addition, performance is further improved by introducing additional features into the ME model and achieves good results in window (-2,+2).

1 Introduction

The task of identifying metaphors for a large-scale corpus has received an increasing amount of attention in the computational linguistics literature. Metaphors, one of figurative languages or tropes, can lead to inaccurate translation in Machine Translation systems and irrelevant document retrieval in Information Retrieval systems. For example, the Chinese word for "翅膀"means literally "wing of an animal". However, when this word appears in a particular context, it has metaphorical expressions. For example,

张开	理想	的	翅膀	(meaning "explore fantasies")
Spread	fantasies	of	wings	

where "翅膀" was not denoted the former literal meaning of "wing", but has a metaphorical expression of "explore fantasies". Information Retrieval systems should exclude this metaphorical expression while searching for "翅膀".

Much research has gone into the processing of metaphors and provides some metaphor understanding systems such as the Met5, which is the first system to recognize examples of metaphors and metonymy under the guidance of preference constraint view4, the Structure-Mapping Engine (SME), a program for studying

[1] Supported by the National Grand Fundamental Research 973 Program of China under Grant No. 2004CB318102; the National High-Tech Research and Development Plan of China under Grant Nos. 2001AA114210，2002AA117010 (863); the National Natural Science Foundation of China Under Grant No.60473138.

Zero Anaphora Resolution in Chinese Discourse

Yuzhen Cui, Qinan Hu, Haihua Pan and Jianhua Hu

Department of Chinese, Translation and Linguistics
City University of Hong Kong, Hong Kong
{50007840@student., qinan.hu@student., cthpan@,
ctjhu@}cityu.edu.hk

Abstract. This paper explores various factors involved in the resolution of zero anaphora in Chinese discourse. Our study differs from previous ones in distinguishing three types of utterances and using clauses as the unit of resolution. The hierarchical structures of utterances enable us to process inter- and intra-utterance anaphora uniformly. Experimental results show that (1) clauses function significantly better than sentences as the unit of resolution, providing an improvement of precision from 36.0% to 63.4%; (2) the inclusion of cataphors and the use of NP forms as a criterion in Cf ranking do not lead to significant improvement of precision; and (3) when assigning antecedents to more than one zero pronoun in the same utterance, the criterion based on grammatical functions gives rise to better performance than that with linear orders.

1 Introduction

Several studies were conducted on zero anaphora for languages like Chinese [1], Japanese [2], Italian [3] and Turkish [4]. The Chinese study resolves zero pronouns in a part-of-speech tagged and shallow-parsed corpus, focusing on pronouns in topic, subject, or object positions in main clauses. All these studies employ Centering Theory (CT) [5, 6] as their framework.

Several problems are found in previous studies. First, it is not clear what counts as an utterance in Chinese discourse. Previous studies either provide no specification or simply use commas and periods as the indicators of utterance ending. Second, the resolution of zero pronouns in subordinate clauses has not been well studied. Third, when two zero pronouns or more occur in the same utterance, it is unclear when they share the same antecedent and when they do not. Finally, cataphora is often not discussed in previous studies.

2 Zero Anaphora in Chinese Discourse

In this study, the term *utterance* refers to an instance of a sentence which is delimited by periods, exclamations, or question marks, and three types of utterances are distinguished, i.e. simple, compound, and complex utterances. A simple utterance consists

Random Walks on Text Structures

Rada Mihalcea

University of North Texas
Computer Science Department
rada@cs.unt.edu

Abstract. Since the early ages of artificial intelligence, associative or semantic networks have been proposed as representations that enable the storage of language units and the relationships that interconnect them, allowing for a variety of inference and reasoning processes, and simulating some of the functionalities of the human mind. The symbolic structures that emerge from these representations correspond naturally to graphs – relational structures capable of encoding the meaning and structure of a cohesive text, following closely the associative or semantic memory representations. The activation or ranking of nodes in such graph structures mimics to some extent the functioning of human memory, and can be turned into a rich source of knowledge useful for several language processing applications. In this paper, we suggest a framework for the application of graph-based ranking algorithms to natural language processing, and illustrate the application of this framework to two traditionally difficult text processing tasks: word sense disambiguation and text summarization.

1 Introduction

Many language processing applications can be modeled by means of a graph. These data structures have the ability to encode in a natural way the meaning and structure of a cohesive text, and follow closely the associative or semantic memory representations. For instance, Figure 1 shows examples of graph representations of textual units[1] and the relationships that interconnect them: 1(a) (adapted from [6]) shows a network of concepts related by semantic relations – simulating a fragment of human memory, on which reasoning and inferences about various concepts represented in the network can be performed; 1(b) shows a network with similar structure, this time automatically derived via definitional links in a dictionary; finally, 1(c) is a graph representation of the cohesive structure of a text, by encoding similarity relationships between textual units.

Provided a graph representation of the text, algorithms for the activation or ranking of nodes in such structures can be used to simulate the functioning of human memory, consequently resulting in solutions for a variety of natural language processing tasks that can be modeled by means of a graph. In this paper, we suggest a framework for the application of graph-based ranking algorithms to text-based graph structures, and show how two text processing applications: word sense disambiguation and text summarization, can find successful solutions within this framework.

[1] We use the term *textual unit* to refer to the textual representation of a *cognitive unit* as defined by Anderson [1]. It can be a word, a concept, a sentence, or any other unit that can find a representation in language.

Shallow Case Role Annotation using Two-Stage Feature-Enhanced String Matching

Samuel W.K. Chan

Dept. of Decision Sciences
The Chinese University of Hong Kong
Hong Kong SAR, China
swkchan@cuhk.edu.hk

Abstract. A two-stage annotation method for identification of case roles in Chinese sentences is proposed. The approach makes use of a feature-enhanced string matching technique which takes full advantage of a huge number of sentence patterns in a Treebank. The first stage of the approach is a coarse-grained syntactic parsing which is complementary to a semantic dissimilarities analysis in its latter stage. The approach goes beyond shallow parsing to a deeper level of case role identification, while preserving robustness, without being bogged down into a complete linguistic analysis. The ideas described have been implemented and an evaluation of 5,000 Chinese sentences is examined in order to justify its significances.

1 Introduction

Automatic information extraction is an area that has received a great deal of attention in recent development of computational linguistics. While a plethora of issues relating to questions of efficiency, flexibility, and portability, amongst others, have been thoroughly discussed, the problem of extracting meaning from natural texts has scarcely been addressed. When the size and quantity of documents available on the Internet are considered, the demand for a highly efficient system that identifies the semantic meaning is clear. Case frame[1], as proposed by most linguists, is one of the most important structures that can be used to represent the meaning of sentences [9]. One could consider a case frame to be a special, or distinguishing, form of knowledge structure about sentences. Although several criteria for recognizing case frames in sentences have been considered in the past, none of the criteria serves as a completely adequate decision procedure. Most of the studies in computational linguistics do not provide any hints on how to map input sentences into case frames automatically, particularly in Chinese. As a result, both the efficiency and robustness of the tech-

[1] Due to the lack of conciseness or conformity that authors have shown in using this and other terms, in this paper, a *case frame* is to be understood as an array of slots, each of which is labelled with a case name, and eventually possibly filled with a case filler, the whole system representing the underlying structure of an input sentence.

SPARTE, a Test Suite for Recognising Textual Entailment in Spanish

Anselmo Peñas, Álvaro Rodrigo, Felisa Verdejo

Dpto. Lenguajes y Sistemas Informáticos, UNED
{anselmo,alvarory,felisa}@lsi.uned.es

Abstract. The aim of Recognising Textual Entailment (RTE) is to determine whether the meaning of a text entails the meaning of another text named hypothesis. RTE systems can be applied to validate the answers of Question Answering (QA) systems. Once the answer to a question is given by the QA system, a hypothesis is built turning the question plus the answer into an affirmative form. If the text (a given document) entails this hypothesis, then the answer is expected to be correct. Thus, a RTE system becomes an Answer Validation system. Within this framework the first problem is to find collections for training and testing RTE systems. We present here the SPARTE corpus aimed at evaluating RTE systems in Spanish. The paper presents the methodology to build SPARTE from the Spanish QA assessments performed at the Cross-Language Evaluation Forum (CLEF) during the last three editions. The paper also describes the test suite and discusses the appropriate evaluation measures together with their baselines.

1 Introduction

The task of Recognising Textual Entailment (RTE) [3] aims at deciding whether the truth of a text entails the truth of another text named hypothesis or, in other words, if the meaning of the hypothesis is enclosed in the meaning of the text. The entailment relation between texts is useful for a variety of tasks as, for example, Automatic Summarisation, where a system could eliminate the passages whose meaning is already entailed by other passages; or Question Answering (QA), where the answer of a question must be entailed by the text that supports the correctness of the answer.

Since RTE task has been defined recently, there exists only few corpora for training and testing RTE systems, and none of them are in Spanish. Thus, we planned the development of SPARTE, a corpus for training and testing RTE systems in Spanish, and specially, systems aimed at validating the correctness of the answers given by QA systems. This automatic Answer Validation would be useful for improving QA systems performance and also for helping humans in the assessment of QA systems output.

SPARTE has been built from the Spanish corpora used at Cross-Language Evaluation Forum (CLEF) for evaluating QA systems during 2003, 2004 and 2005. At the end of development, SPARTE contains 2962 hypothesis with a document label and a TRUE/FALSE value indicating whether the document entails the hypothesis or not.

Section 2 describes the development of SPARTE in detail. Section 3 evaluates some features of the corpus. Section 4 discusses and suggests the way of using SPARTE for evaluation purposes. Section 5 is devoted to some other corpora related to RTE. Finally, we give some conclusions and future work.

Analysis of a Textual Entailer

Vasile Rus[1], Philip M. McCarthy[2], and Arthur C. Graesser[2]

[1] Department of Computer Science
[2] Department of Psychology
Institute for Intelligent Systems
The University of Memphis
Memphis, TN 38120, USA
{vrus, pmmccrth, a-graesser}@memphis.edu

Abstract. We present in this paper the structure of a textual entailer, offer a detailed view of lexical aspects of entailment and study the impact of syntactic information on the overall performance of the textual entailer. It is shown that lemmatization has a big impact on the lexical component of our approach and that syntax leads to accurate entailment decisions for a subset of the test data.

1 Introduction

The task of textual entailment is to decide whether a text fragment the size of a sentence, called the Text (T), can logically infer another text of same or smaller size, called the Hypothesis (H).

Entailment has received a great deal of attention since it was proposed (in 2004) under the Recognizing Textual Entailment (RTE) Challenge [7]. In our experiments presented here, we use the standard data set that RTE offers for development and comparison purposes.

The purpose of this paper is to perform an analysis of the textual entailer presented in [8]. In particular, we consider the three main subsystems of the entailer: the lexical component, the syntactic component and the negation handling component. We study each element's contribution to the performance of the system or a part of it. Different aspects of entailment have been analyzed by different groups. The Related Work section describes previous work on entailment analysis. Here, we analyze the task from a systemic, component angle. For instance, we report the impact of lemmatization for entailment, which, as far as we are aware, has yet to be reported. This type of analysis is important to better understand the interaction among different processing modules and to improve decisions as to whether the inclusion of a particular component is advantageous.

In our study, we conduct two levels of analysis. First, we look at how a particular feature impacts the component to which it belongs. For instance, lemmatization is part of the lexical component and we report how the lexical score changes depending upon its presence. Second, we present how a particular feature affects the overall entailment performance. The reader should note that our solution to entailment is based on a limited number of external resources and thus components such as world knowledge are not investigated: we use lexical

Referring via Document Parts

Ivandré Paraboni[1] and Kees van Deemter[2]

[1] Instituto de Ciências Matemáticas e de Computação - ICMC
Universidade de São Paulo - USP
Av Trabalhador São-Carlense, 400, 13560-970 - São Carlos SP, Brazil
ivandre@icmc.usp.br

[2] Department of Computing Science, King's College, University of Aberdeen
Aberdeen AB24 3UE, Scotland, UK
kvdeemte@csd.abdn.ac.uk

Abstract. Documents in a wide range of genres often contain references to their own sections, pictures etc. We call such referring expressions instances of *Document Deixis*. The present work focuses on the generation of Document Deixis in the context of a particular kind of natural language generation system in which these descriptions are not specified as part of the input, i.e., when it is up *to the system* to decide whether a reference is called for and, if so, which document entity it should refer to. We ask under what circumstances it is advantageous to describe domain objects in terms of the document parts where they are mentioned (as in "the insulin described in section 2"). We report on an experiment suggesting that such indirect descriptions are preferred by human readers whenever they cause the generated descriptions to be shorter than they would otherwise be.

1 Introduction

Document parts such as sections, subsections, pictures, paragraphs etc may be referred to for various purposes, for example to point to additional information on the current topic of the text, e.g., "see also section 7". References to document parts will often be *deictic*, in the sense that the realisation of the expression depends on the place in the document where the referring expression is uttered (e.g., "this section" versus "section 1.1."). Accordingly, we will call the references to parts of the *same* document instances of Document Deixis (DDX).

We are interested in a particular kind of DDX, which we have previously called *object-level* instances of Document Deixis [9]. These are usually part of a larger expression which refers to a domain entity. The entity in question may be concrete (e.g., the medicines in Example 1) or abstract (e.g., the advice in Example 2). In the corpora that we investigated – patient information leaflets [1] - references to abstract entities or sets of them are far more common.

Generation of Natural Language Explanations of Rules in an Expert System*

María de los Ángeles Alonso-Lavernia,[1]
Argelio Víctor De-la-Cruz-Rivera,[2] and Grigori Sidorov [1]

[1] Center for Computing Research (CIC), National Polytechnic Institute (IPN),
Av. Juan de Dios Bátiz, Zacatenco, DF, 07738, Mexico
marial@uaeh.uaehred.mx, sidorov@cic.ipn.mx

[2] Center for Research on Technologies of Information and Systems (CITIS),
Autonomous University of Hidalgo State (UAEH), Mexico

Abstract. We present a domain-independent method for generation of natural language explanations of rules in expert systems. The method is based on explanatory rules written in a procedural formal language, which build the explanation from predefined natural language texts fragments. For better style, a specific text fragment is randomly selected from a group of synonymous expressions. We have implemented 16 groups of explanatory rules and 74 groups of explanatory texts containing about 200 text fragments.

1 Introduction

Expert systems are widely used to solve particular problems in a rather narrow area of expertise. They are based on knowledge obtained during interaction with human experts in the field, so they are also often referred to as *knowledge-based systems*.

One of important requirements for an expert system is the system's ability to explain its conclusions in a manner understandable to the user. The best form of presenting such an explanation is a text in natural language [5]. One approach to generation of explanations is to use as explanation the rules from the knowledge base that were fired during reasoning [6]. Another approach is writing special code that paraphrases the rules [8]. These approaches do not allow for description of the ideas behind the fired rules. An alternative is to use another knowledge database for generation of explanations [7]. This approach requires a double amount of work for constructing knowledge bases.

In this paper, we present a method that allows for representation of the ideas behind the rules, does not require any additional knowledge bases, and is domain-independent—i.e., it does not require reprogramming of an explanation system if the knowledge base is changed.

* This work was done under partial support of Mexican Government (CONACyT, SNI), Autonomous University of Hidalgo State, and National Polytechnic Institute, Mexico.

Generating a Set of Rules to Determine Honorific Expression Using Decision Tree Learning

Kanako Komiya, Yasuhiro Tajima, Nobuo Inui and Yoshiyuki Kotani

Department of Computer, Information and Communication Sciences
Tokyo University of Agriculture and Technology
2-24-16, Nakacho, Koganei, Tokyo, Japan, 184-8588
komiya@fairy.ei.tuat.ac.jp

Abstract. In Japanese language, the speaker must choose suitable honorific expressions depending on many factors. The computer system should imitate this mechanism to make a natural Japanese sentence. We made a system to determine a suitable expression and named it honorific expression determining system (HEDS). It generates a set of rules to determine suitable honorific expression automatically, by decision tree learning. The system HEDS determines one out of the three classes for an input sentence: the respect expression, the modesty expression and the non-honorific expression and determines what expression the verb is. We calculated the accuracy of HEDS using the cross validation method and it was up to 74.88%.

1 Introduction

In Japanese language, one must choose suitable honorific expressions depending on the speaker, the addressees, the subject of the utterance, contents of the dialogue and situations in the conversation. The computer system should imitate this mechanism to make a natural Japanese sentence.

Japanese language has the two types of honorific expression: (1) respect or modesty expression and (2) polite expression. The respect expression is used to display respect to others, or their higher rank, and practically to show second person of the sentential implicit subject, in contrast that the modesty expression shows first person. The modesty expression is an expression that one display modesty to respecting persons. These two honorific expressions cannot be used in a single word at the same time, but the combination of (1) and (2) can be used for one word at the same time. We focus on the type (1) in this paper.

2 Honorific Expression Determining System (HEDS)

The user of HEDS provides honorific expressions and its factors for determining suitable honorific expressions as data. Then HEDS generates a set of selection rules. It

NLP (Natural Language Processing) for NLP (Natural Language Programming)

Rada Mihalcea[1], Hugo Liu[2], Henry Lieberman[2]

[1] Computer Science Department, University of North Texas
rada@cs.unt.edu
[2] Media Arts and Sciences, Massachusetts Institute of Technology
{hugo,henry}@media.mit.edu

Abstract. Natural Language Processing holds great promise for making computer interfaces that are easier to use for people, since people will (hopefully) be able to talk to the computer in their own language, rather than learn a specialized language of computer commands. For programming, however, the necessity of a formal programming language for communicating with a computer has always been taken for granted. We would like to challenge this assumption. We believe that modern Natural Language Processing techniques can make possible the use of natural language to (at least partially) express programming ideas, thus drastically increasing the accessibility of programming to non-expert users. To demonstrate the feasibility of Natural Language Programming, this paper tackles what are perceived to be some of the hardest cases: steps and loops. We look at a corpus of English descriptions used as programming assignments, and develop some techniques for mapping linguistic constructs onto program structures, which we refer to as programmatic semantics.

1 Introduction

Natural Language Processing and Programming Languages are both established areas in the field of Computer Science, each of them with a long research tradition. Although they are both centered around a common theme – "languages" – over the years, there has been only little interaction (if any) between them[3]. This paper tries to address this gap by proposing a system that attempts to convert natural language text into computer programs. While we overview the features of a natural language programming system that attempts to tackle both the descriptive and procedural programming paradigms, in this paper we focus on the aspects related to procedural programming. Starting with an English text, we show how a natural language programming system can automatically identify steps, loops, and comments, and convert them into a program *skeleton* that can be used as a starting point for writing a computer program, expected to be particularly useful for those who begin learning how to program.

We start by overviewing the main features of a descriptive natural language programming system METAFOR introduced in recent related work [6]. We then describe in detail the main components of a procedural programming system as introduced in this

[3] Here, the obvious use of programming languages for coding natural language processing systems is not considered as a "meaningful" interaction.

Balancing transactions in practical dialogues

Luis Pineda, Hayde Castellanos, Sergio Coria, Varinia Estrada, Fernanda López,
Isabel López, Ivan Meza, Iván Moreno, Patricia Pérez, Carlos Rodríguez

Instituto de Investigaciones en Matemáticas Aplicadas y en Sistemas (IIMAS)
Universidad Nacional Autónoma de México (UNAM)
Cto. Escolar S/N, Cd. Universitaria, Coyoacán, México, D. F.
luis@leibniz.iimas.unam.mx

Abstract. In this paper a theory of dialogue acts analysis in problem-solving tasks-oriented conversations is presented. The theory postulates that in practical dialogues every transaction has a component in the obligations and the common ground planes of expression, and contributions made by dialogue acts making a "charge" in the transaction should be "balanced" by contributions making the corresponding "credit", and a complete transaction is balanced in both of these planes. In addition, transactions have a structure which constraints strongly the realization of dialogue acts. A dialogue act tagging methodology based on the theory is also presented. The theory and its related methodology have been applied to the analysis of a multimodal corpus in a design task, and the figures of the agreement reached in the preliminary experiments are presented.

1 Introduction

In this paper a theory for the analysis of dialog acts in practical dialogs is presented. In this theory dialogues acts are analyzed in relation to the obligations and common ground structures of task oriented conversations, and we provide an explicit analysis and tagging methodology for these two dialogue structures. According to Allen et al. [1], practical dialogues have the purpose to achieve a concrete goal, and the conversational competence required to engage in this kind of dialogs is significantly simpler than general human conversation (i.e. the practical dialogue hypothesis) and the main aspects of language interpretation and dialogue management are domain independent (i.e. domain independence hypothesis). Simple dialogues can be reduced to achieve a single goal and involve only one transaction, but often the dialogue involves a sequence of transactions. From the empirical study of a corpus in the kitchen design domain we suggest that transactions are also characterized in terms of an intention specification phase, followed by the intention satisfaction phase, and the structure of the dialogue is closely related to the structure of the problem-solving task, and in this regard, our approach loosely resembles Grosz and Sidner's discourse theory [7]. We also postulate the hypothesis that transactions can be analyzed in terms of their conversational obligations and common ground structures, and that complete transactions

Finite State Grammar Transduction from Distributed Collected Knowledge

Rakesh Gupta[1] and Ken Hennacy[2]

[1] Honda Research Institute USA, Inc.
800 California Street, Suite 300
Mountain View, CA 94041
rgupta@hra.com
[2] University of Maryland
Institute for Advanced Computer Studies
College Park, MD 20742
khennacy@cs.umd.edu

Abstract. In this paper, we discuss the use of Open Mind Indoor Common Sense (OMICS) project for the purpose of speech recognition of user requests. As part of OMICS data collection, we asked users to enter different ways of asking a robot to perform specific tasks. This paraphrasing data is processed using Natural Language techniques and lexical resources like WordNet to generate a Finite State Grammar Transducer (FSGT). This transducer captures the variations in user requests and captures their structure.
We compare the task recognition performance of this FSGT model with an n-gram Statistical Language Model (SLM). The SLM model is trained with the same data that was used to generate the FSGT. The FSGT model and SLM are combined in a two-pass system to optimize full and partial recognition for both in-grammar and out-of-grammar user requests. Our work validates the use of a web based knowledge capture system to harvest phrases to build grammar models. Work was performed using Nuance Speech Recognition system.

1 Introduction

Humans often wish to communicate with robots about what they like done. It is awkward to be constrained to specific set of commands. Therefore, a free-form interface that supports natural human robot interaction is desirable.

A finite state transducer is a finite automaton whose state transitions are labeled with both input and output labels. A path through the transducer encodes a mapping from an input symbol sequence to an output symbol sequence [1]. Grammar is a structure that defines a set of phrases that a person is expected to say. In this work, our goal is to automate the process of creating a Finite State Grammar Transducer (FSGT) to map utterances to task labels from text data contributed by volunteers over the web.

It is a challenge to develop a grammar that will recognize a large variety of phrases and achieve a high recognition accuracy. Manual creation of a set of

Predicting Dialogue Acts from Prosodic Information

Sergio Coria, Luis Pineda

Institute for Applied Mathematics and Systems (IIMAS)
Universidad Nacional Autonoma de Mexico (UNAM)
Circuito Escolar S/N, Ciudad Universitaria, Del. Coyoacan
04510 Mexico, D.F. Mexico
coria@turing.iimas.unam.mx, luis@leibniz.iimas.unam.mx

Abstract. In this paper, the influence of intonation to recognize dialogue acts from speech is assessed. Assessment is based on an empirical approach: manually tagged data from a spoken-dialogue and video corpus are used in a CART-style machine learning algorithm to produce a predictive model. Our approach involves two general stages: the tagging task, and the development of machine learning experiments. In the first stage, human annotators produce dialogue act taggings using a formal methodology, obtaining a highly enough tagging agreement, measured with Kappa statistics. In the second stage, tagging data are used to generate decision trees. Preliminary results show that intonation information is useful to recognize sentence mood, and sentence mood and utterance duration data contribute to recognize dialogue act. Precision, recall and Kappa values of the predictive model are promising. Our model can contribute to improve automatic speech recognition or dialogue management systems.

1 Introduction

A dialogue act tag characterizes the type of intention which a speaker intends to express in an utterance. A listener has to analyze the utterance, its intonation and its context to identify the correct dialogue act which his interlocutor wants to communicate. Two models to analyze dialogue acts are DAMSL (Dialogue Act Markup in Several Layers) [1] and DIME-DAMSL [2]; the latter is a multimodal adaptation of DAMSL to the DIME project [3]. The Verbmobil Project [4] developed another dialogue act model, which has been used in practical dialogue systems.

DAMSL assumes that dialogue acts occur on four dimensions: communicative status, information level, forward and backward looking function. The communicative status determines if an utterance was uninterpretable or abandoned or if it expressed a self-talk. The information level classifies utterances according to whether they refer to the task, the task management, or the communication management. The forward looking function identifies the effect which an utterance has on the future of the dialogue; this includes statements (assert, reassert), influencing an addressee future actions (open option, action directive), information requests, commiting a speaker future actions (offer, commit), conventional (opening, closing), explicit performative, or exclamation. Backward looking function indicates the way an utterance relates to one or more previous utterances; this includes agreement (accept, accept part, maybe, reject

Disambiguation Based on Wordnet for Transliteration of Arabic Numerals for Korean TTS

Youngim Jung[1], Aesun Yoon[2], and Hyuk-Chul Kwon[1]

[1]Pusan National University, Department of Computer Science and Engineering,
Jangjeon-dong Geumjeong-gu, 609-735 Busan, S. Korea
{acorn, hckwon}@pusan.ac.kr
[2]Pusan National University, Department of French,
Jangjeon-dong Geumjeong-gu, 609-735 Busan, S. Korea
asyoon@pusan.ac.kr

Abstract Transliteration of Arabic numerals is not easily resolved. Arabic numerals occur frequently in scientific and informative texts and deliver significant meanings. Since readings of Arabic numerals depend largely on their context, generating accurate pronunciation of Arabic numerals is one of the critical criteria in evaluating TTS systems. In this paper, (1) contextual, pattern, and arithmetic features are extracted from a transliterated corpus; (2) ambiguities of homographic classifiers are resolved based on the semantic relations in KorLex1.0 (Korean Lexico-Semantic Network); (3) a classification model for accurate and efficient transliteration of Arabic numerals is proposed in order to improve Korean TTS systems. The proposed model yields 97.3% accuracy, which is 9.5% higher than that of a customized Korean TTS system.

1 Introduction

TTS technologies for naturalness have improved dramatically and have been applied to many unlimited systems in terms of domain. However, improvement on the technique for accurate transliteration of non-alphabetic symbols such as Arabic numerals and various text symbols[1] has been relatively static.

According to the accuracy test results of 19 TTS products by Voice Information Associates, the weakest area of TTS products is in number processing in the following ambiguity-generating areas, as shown in Table 1 [10].

Table 1. TTS Accuracy Test Results Summary

Test area	Accuracy (%)
Number	**55.6**
Word of Foreign Origin	58.8
Acronym	74.1

[1] Since Arabic numerals and text symbols have graphic simplicity and deliver more precise information, the occurrence of Arabic numerals and text symbols is as high as 8.31% in Korean newspaper articles.

MFCRank: A Web Ranking Algorithm Based on Correlation of Multiple Features

Yunming Ye[1], Yan Li[1], Xiaofei Xu[1], Joshua Huang[2], and Xiaojun Chen[1]

[1] Shenzhen Graduate School, Harbin Institute of Technology, Shenzhen 518055, China
yym_sjtu@yahoo.com.cn
[2] E-Business Technology Institute, The University of Hong Kong, Hong Kong
jhuang@eti.hku.hk

Abstract. This paper presents a new ranking algorithm ***MFCRank*** for topic-specific Web search systems. The basic idea is to correlate two types of similarity information into a unified link analysis model so that the rich content and link features in Web collections can be exploited efficiently to improve the ranking performance. First, a new surfer model ***JBC*** is proposed, under which the topic similarity information among neighborhood pages is used to weigh the jumping probability of the surfer and to direct the surfing activities. Secondly, as *JBC* surfer model is still query-independent, a correlation between the query and *JBC* is essential. This is implemented by the definition of *MFCRank* score, which is the linear combination of *JBC* score and the similarity value between the query and the matched pages. Through the two correlation steps, the features contained in the plain text, link structure, anchor text and user query can be smoothly correlated in one single ranking model. Ranking experiments have been carried out on a set of topic-specific Web page collections. Experimental results showed that our algorithm gained great improvement with regard to the ranking precision.

Keywords: Ranking, Search Engine, Link Analysis, *PageRank*, Web

1 Introduction

The enormous volume of the Web presents a big challenge to Web search, as there are always too many results returned for specific queries, and going through the entire results to find the desired information is very time-consuming for the user. To improve the information retrieval efficiency, Web search engines need to employ a suitable page ranking strategy to correctly rank the search results so that the most relevant (or important) pages will be included in the top list of the search results.

In traditional information retrieval, ranking measures, such as *TF*IDF* [1], usually rely on the text features alone to rate plain text documents. This strategy can give poor results on the Web, due to the fact that the indexed Web

On Text Ranking for Information Retrieval based on Degree of Preference

Bo-Yeong Kang[1] and Dae-Won Kim[2,⋆]

[1]Center of Healthcare Ontology R&D, Seoul National University
Yeongeon-dong, Jongro-gu, Seoul, Korea
[2]School of Computer Science and Engineering, Chung-Ang University
221 Heukseok-dong, Dongjak-gu, Seoul, Korea
(⋆Corresponding author: dwkim@cau.ac.kr)

Abstract. A great deal of research has been made to model the vagueness and uncertainty in information retrieval. One such research is fuzzy ranking models, which have been showing their superior performance in handling the uncertainty involved in the retrieval process. However, these conventional fuzzy ranking models are limited to incorporate the user preference when calculating the rank of documents. To address this issue, we develop a new fuzzy ranking model based on the user preference.

1 Introduction

In recent years a great deal of research in information retrieval has aimed at modelling the vagueness and uncertainty which invariably characterize the management of information. The application of fuzzy set theory to IR have concerned the representation of documents and the query [1], and many fuzzy ranking models such as MMM, PAICE, and P-NORM have been showing their superior performance in handling the uncertainty in the retrieval process [2–4]. The ranking is achieved by calculating a similarity between two fuzzy sets, a document D and a query Q. However, in spite that the user has an ability to reflect their preference for the information need in searching, these conventional models are limited to incorporate the user preference when calculating the rank of documents. Let us suppose that we are given a vector of query Q with a fuzzy set of the term and its membership degree:

$Q = \{fuzzy(0.8), IR(0.7), korea(0.3), author(0.2)\}$

A document collection consists of four documents (D_1, D_2, D_3, D_4) in which each document is represented as a fuzzy set of the index term and its weight.

$D_1 = \{fuzzy(0.8), IR(0.7)\}$
$D_2 = \{fuzzy(0.2), IR(0.2), korea(0.3), author(0.2)\}$
$D_3 = \{korea(0.7), IR(0.8)\}$
$D_4 = \{fuzzy(0.8), IR(0.7), korea(0.3), author(0.2)\}$

Given a query Q, we are wondering what is the best result of ranking? Intuitively, we know that D_4 is the most relevant document and D_3 is the least

Lexical Normalization and Relationship Alternatives for a Term Dependence Model in Information Retrieval

Marco Gonzalez[1], Vera L. S de Lima[1], and José V. de Lima[2]

[1] PUCRS - Faculdade de Informática
Av. Ipiranga, 6681 – Prédio 16 - PPGCC
90619-900 Porto Alegre, Brazil
{gonzalez, vera} @inf.pucrs.br
[2] UFRGS – Instituto de Informática
Av. Bento Gonçalves, 9500
91501-970 Porto Alegre, Brazil
valdeni@inf.ufrgs.br

Abstract. We analyze alternative strategies for lexical normalization and term relationship identification for a dependence structured indexing system [14], in the probabilistic retrieval approach. This system uses a dependence parse tree and Chow expansion [5]. Stemming, lemmatizing, and nominalization processes are tested as lexical normalization, while head-modifier pairs and binary lexical relations are tested as term relationships. We demonstrate that our proposal, binary lexical relations with nominalized terms for Portuguese, contributes to the performance improvement in information retrieval.

1 Introduction

Many information retrieval (IR) systems are based on the assumption that each term is statistically independent of all other terms in the text. Those systems have been developed because this independence leads to a formal representation of the probabilistic approach more easily. But, the independence assumption is understood to be inconsistent [6] and there are regularities provided by term dependences that need to be considered [16].

Some models have been proposed to incorporate term dependence strategies (e.g., [19], [16]). However, the formal representation of the probabilistic approach cannot be easily maintained when there are no constraints for term relationships, i.e., when a higher order model of term dependence is applied. For reducing this problem, Rijsbergen [19] adopted the algorithm proposed by Chow and Liu [5] that uses a maximum spanning tree for incorporating term dependence into a probabilistic approach.

Adapting the Rijsbergen's strategy, Changki Lee and Gary Lee [14] presented a method for incorporating term dependence into the probabilistic retrieval approach using Chow expansion. They proposed a dependence structured indexing (DSI) sys-

Web Search Model for Dynamic and Fuzzy Directory Search

Bumghi Choi[1], Ju-Hong Lee[1], Sun Park[2], Tae-Su Park[2]

School of Computer Science and Engineering, Inha University, Incheon, Korea
[1] { neural, juhong}@inha.ac.kr
[2]{sunpark,taesu}@datamining.inha.ac.kr

Abstract. In web search engines, index search used to be evaluated at a high recall rate. However, the pitfall is that users have to work hard to select relevant documents from too many search results. Skillful surfers tend to prefer the index searching method, while on the other hand, those who are not accustomed to web searching generally use the directory search method. Therefore, the directory searching method is needed as a complementary way of web searching. However, in the case that target documents for searching are obscurely categorized or users have no exact knowledge about the appropriate categories of target documents, occasionally directory search will fail to come up with satisfactory results. That is, the directory search method has a high precision and low recall rate. With this motive, we propose a novel model in which a category hierarchy is dynamically constructed. To do this, a category is regarded as a fuzzy set which includes keywords. Similarly extensible subcategories of a category can be found using fuzzy relational products. The merit of this method is to enhance the recall rate of directory search by reconstructing subcategories on the basis of similarity.

1 Introduction

The index searching method has an advantage in that it quickly searches the documents indexed by an input keyword. However, it may exhibit a critical defect by generating too many results or failing to search even a single one of the targeted documents. It is because that given keywords can't be satisfactorily matched with the subjects of the target documents, or they happen to be heteronyms or homonyms, or the target documents may not be properly indexed by the keywords inside them.

In spite of many advantages of the index searching method and many efforts to improve its efficiency, we absolutely need the directory search as a complementary method of the index search. Especially for beginners, the directory searching method is preferred because it can zoom in more detailed subjects by reconstructing the subcategories of a category in a fast manner if they are familiar with the exact information of the categorization of the search subjects. However, if users don't know the categories regarding the subjects of the target documents, or if documents are not exactly categorized, it can't provide users with satisfactory results, and occasionally it causes inconvenience by navigating too many categories before reaching the targets[4, 5, 6].

Information Retrieval from Spoken Documents [⋆]

Michal Fapšo, Pavel Smrž, Petr Schwarz, Igor Szöke, Milan Schwarz,
Jan Černocký, Martin Karafiát, and Lukáš Burget

Faculty of Information Technology, Brno University of Technology,
Božetěchova 2, 612 66 Brno, Czech Republic
speech@fit.vutbr.cz, http://www.fit.vutbr.cz/speech/

Abstract. This paper describes a designed and implemented system for efficient storage, indexing and search in collections of spoken documents that takes advantage of automatic speech recognition. As the quality of current speech recognizers is not sufficient for a great deal of applications, it is necessary to index the ambiguous output of the recognition, i. e. the acyclic graphs of word hypotheses — recognition lattices. Then, it is not possible to directly apply the standard methods known from text-based systems. The paper discusses an optimized indexing system for efficient search in the complex and large data structure that has been developed by our group. The search engine works as a server. The meeting browser JFerret, developed withing the European AMI project, is used as a client to browse search results.

1 Introduction

The most straightforward way to use a large vocabulary continuous speech recognizer (LVCSR) to search in audio data is to use existing search engines on the textual ("1-best") output from the recognizer. For such data, it is possible to use common text indexing techniques. However, these systems have satisfactory results only for high quality speech data with correct pronunciation. In the case of low quality speech data (noisy TV and radio broadcast, meetings, teleconferences) it is highly probable that the recognizer scores a word which is really in the speech worse than another word.

We can however use a richer output of the recognizer – most recognition engines are able to produce an oriented graph of hypotheses (called *lattice*). On contrary to 1-best output, the lattices tend to be complex and large. For efficient searching in such a complex and large data structure, the creation of an optimized indexing system which is the core of each fast search engine is necessary. The proposed system is based on principles used in Google [1]. It consists of **indexer**, **sorter** and **searcher**.

[⋆] This work was partly supported by European project AMI (Augmented Multiparty Interaction, FP6-506811) and Grant Agency of Czech Republic under project No. 102/05/0278. Pavel Smrž was supported by MŠMT Research Plan MSM 6383917201. The hardware used in this work was partially provided by CESNET under project No. 119/2004.

Automatic Image Annotation based on WordNet and Hierarchical Ensembles

Wei Li and Maosong Sun

State Key Lab of Intelligent Technology and Systems
Department of Computer Science and Technology, Tsinghua University
Beijing 100084, China
wei.lee04@gmail.com, sms@mail.tsinghua.edu.cn

Abstract. Automatic image annotation concerns a process of automatically labeling image contents with a pre-defined set of keywords, which are regarded as descriptors of image high-level semantics, so as to enable semantic image retrieval via keywords. A serious problem in this task is the unsatisfactory annotation performance due to the semantic gap between the visual content and keywords. Targeting at this problem, we present a new approach that tries to incorporate lexical semantics into the image annotation process. In the phase of training, given a training set of images labeled with keywords, a basic visual vocabulary consisting of visual terms, extracted from the image to represent its content, and the associated keywords is generated at first, using K-means clustering combined with semantic constraints obtained from WordNet, then the statistical correlation between visual terms and keywords is modeled by a two-level hierarchical ensemble model composed of probabilistic SVM classifiers and a co-occurrence language model. In the phase of annotation, given an unlabeled image, the most likely associated keywords are predicted by the posterior probability of each keyword given each visual term at the first-level classifier ensemble, then the second-level language model is used to refine the annotation quality by word co-occurrence statistics derived from the annotated keywords in the training set of images. We carried out experiments on a medium-sized image collection from Corel Stock Photo CDs. The experimental results demonstrated that the annotation performance of this method outperforms some traditional annotation methods by about 7% in average precision, showing the feasibility and effectiveness of the proposed approach.

1 Introduction

With the exponential growth of multimedia information, efficient access to a large image/video databases is highly desired. To address this problem, Content-Based Visual Information Retrieval, has become a hot research topic in the domain of both computer vision and information retrieval in the last decade.

Traditionally, most of the content-based image retrieval techniques is based on the query-by-example (QBE) architecture, in which user should provide an image example firstly, the visual similarity of low-level visual features such as color, texture and shape descriptors is then computed to find the visually similar images compared to

Creating a Testbed for the Evaluation of Automatically Generated Back-of-the-book Indexes

Andras Csomai and Rada Mihalcea

University of North Texas
Computer Science Department
csomaia@unt.edu, rada@cs.unt.edu

Abstract. The automatic generation of back-of-the book indexes seems to be out of sight of the Information Retrieval and Natural Language Processing communities, although the increasingly large number of books available in electronic format, as well as recent advances in keyphrase extraction, should motivate an increased interest in this topic. In this paper, we describe the background relevant to the process of creating back-of-the-book indexes, namely (1) a short overview of the origin and structure of back-of-the-book indexes, and (2) the correspondence that can be established between techniques for automatic index construction and keyphrase extraction. Since the development of any automatic system requires in the first place an evaluation testbed, we describe our work in building a gold standard collection of books and indexes, and we present several metrics that can be used for the evaluation of automatically generated indexes against the gold standard. Finally, we investigate the properties of the gold standard index, such as index size, length of index entries, and upper bounds on coverage as indicated by the presence of index entries in the document.

1 Introduction

> *"Knowledge is of two kinds. We know a subject ourselves, or we know where we can find information on it." (Samuel Johnson)*

The automatic construction of back-of-the-book indexes is one of the few tasks related to publishing that still requires extensive human labor. While there is a certain degree of computer assistance, mainly consisting of tools that help the professional indexer organize and edit the index, there are however no methods or tools that would allow for a complete or nearly-complete automation. Despite the lack of automation in this task, there is however another closely related natural language processing task – kepyphrase extraction – where in recent years we have witnessed considerable improvements.

In this paper, we argue that the task of automatic index construction should be reconsidered in the light of the progress made in the task of keyphrase extraction. We show how, following methodologies used for the evaluation of keyphrase extraction systems, we can devise an evaluation methodology for back-of-the-book indexes, including a gold standard dataset and a set of evaluation metrics.

Automatic acquisition of semantic-based question reformulations for question answering

Jamileh Yousefi and Leila Kosseim

CLaC laboratory
Department of Computer Science and Software Engineering
1400 de Maisonneuve Blvd. West
Montreal, Quebec, Canada H3G 1M8
j_yousef@cs.concordia.ca, kosseim@cs.concordia.ca

Abstract. In this paper, we present a method for the automatic acquisition of semantic-based reformulations from natural language questions. Our goal is to find useful and generic reformulation patterns, which can be used in our question answering system to find better candidate answers. We used 1343 examples of different types of questions and their corresponding answers from the TREC-8, TREC-9 and TREC-10 collection as training set. The system automatically extracts patterns from sentences retrieved from the Web based on syntactic tags and the semantic relations holding between the main arguments of the question and answer as defined in WordNet. Each extracted pattern is then assigned a weight according to its length, the distance between keywords, the answer sub-phrase score, and the level of semantic similarity between the extracted sentence and the question. The system differs from most other reformulation learning systems in its emphasis on semantic features. To evaluate the generated patterns, we used our own Web QA system and compared its results with manually created patterns and automatically generated ones. The evaluation on about 500 questions from TREC-11 shows comparable results in precision and MRR scores. Hence, no loss of quality was experienced, but no manual work is now necessary.

1 Introduction

Question reformulation deals with identifying possible forms of expressing answers given a natural language question. These reformulations can be used in a QA system to retrieve answers in a large document collection. For example given the question *What is another name for the North Star?*, a reformulation-based QA system will search for formulations like *<NP>, another name for the North Star* or *<NP> is another name for the North Star* in the document collection and will instantiate <NP> with the matching noun phrase. The ideal reformulation should not retrieve incorrect answers but should also identify many candidate answers.

Using N-gram Models to Combine Query Translations in Cross-Language Question Answering

Rita M. Aceves-Pérez, Luis Villaseñor-Pineda, Manuel Montes-y-Gómez

Language Technologies Group, Computer Science Department,
National Institute of Astrophysics, Optics and Electronics (INAOE), Mexico.
{rmaceves, mmontesg, villasen}@inaoep.mx

Abstract. This paper presents a method for cross-language question answering. The method combines multiple query translations in order to improve the answering precision. The combination of translations is based on their pertinence to the target document collection rather than on their grammatical correctness. The pertinence is measured by the translation perplexity with respect to the collection language model. Experimental evaluation on question answering demonstrates that the proposed approach outperforms the results obtained by the best translation machine.

1 Introduction

A question answering (QA) system is a particular kind of search engine that allows users to ask questions using natural language instead of an artificial query language. In a cross-lingual scenario the questions are formulated in a language different from the document collection. In this case, the efficiency of the QA system greatly depends on the way it confronts the idiomatic barrier. Traditional approaches for cross-lingual information access involve translating either the documents into the expected query language or the questions into the document language. The first approach is not always practical, in particular when the document collection is very large. The second approach is more common. However, because of the small size of questions in QA, the machine translation methods do not have enough context information, and tend to produce unsatisfactory question translations.

A bad question translation generates a cascade error through all phases of the QA process. This effect is evident in the results of cross-lingual QA reported on the last edition of CLEF [4]. For instance, the results from the best cross-lingual system (that uses the French as target language) were 64% of precision for the monolingual task, and 39.5% when using English as question language. In this case, the errors in the translation of the question cause a drop in precision of 61.7%.

Recent methods for cross-lingual information access attempt to minimize the error introduced by the translation machines. In particular, the idea of combining the capacities of several translation machines has been successfully used in cross-lingual information retrieval [2]. In this field, most works focus on the selection of the best

[*] This work was partially financed by the CONACYT (grants 43990 and 184663). We also like to thanks to the CLEF for the provided resources.

A Question Answering System on Special Domain and the Implementation of Speech Interface

Haiqing Hu[1,2], Fuji Ren[1], Shingo Kuroiwa[1], and Shuwu Zhang[3]

[1] Faculty of Engineering, The University of Tokushima,
Tokushimashi 770-8506, Japan,
{huhq, ren, kuroiwa}@is.tokushima-u.ac.jp
[2] Xian University of Technology, Xian, China
[3] The Institute of Automation Chinese Academy of Sciences, Beijing, China
swzhang@hitic.ia.ac.cn

Abstract. In this paper, we propose a construction of Question Answering(QA) system, which synthesizes the answers retrieval from the frequent asked questions database and documents database, based on special domain about sightseeing information. A speech interface for the special domain was implemented along with the text interface, using an acoustic model HMM, a pronunciation lexicon, and a language model FSN on the basis of the feature of Chinese sentence patterns. We consider the synthetic model based on statistic VSM and shallow language analysis for sightseeing information. Experimental results showed high accuracy can be achieved for the special domain and the speech interface is available for frequently asked questions about sightseeing information.

Keywords : Question Answering System, Similarity Computing, Special Domain, FSN, Speech Recognition, Chinese

1 Introduction

Question Answering (QA) is a technology that aims at retrieving the answer of a question written in natural language in large collections of documents. QA systems are presented with natural language questions and the expected output is either the exact answer identified in a text or small text fragments containing the answer. A lot of research has been done on the QA technology, and the technology relates to a lot of fields of NLP (Natural Language Processing), such as Information Retrieval(IR), Information Extraction(IE), Conversation Interface, etc. Recently, systems based on statistical retrieval techniques and shallow language analysis are much used techniques in answer retrieval using natural language. In the Question Answering task of TREC(Text REtrieval Conference) QA track, the target has become the open domain (the search object domain to the questions is not limited) in recent years. But the treatment of special domains and the construction of a practical QA system as a specialist are very difficult. On the other hand, it is easier to use special domain knowledge by

Multi-Document Summarization Based on BE-Vector Clustering

Dexi Liu[1,2,3], Yanxiang He[1,3], Donghong Ji[3,4], and Hua Yang[1,3]

[1] School of Computer, Wuhan University, Wuhan 430079, P. R. China
[2] School of Physics, Xiangfan University, Xiangfan 441053, P. R. China
[3] Center for Study of Language and Information, Wuhan University,
Wuhan 430079, P. R. China
[4] Institute for Infocomm Research, Heng Mui Keng Terrace 119613, Singapore
dexiliu@gmail.com, yxhe@whu.edu.cn,
dhji@i2r.a-star.edu.sg, yh@eis.whu.edu.cn

Abstract. In this paper, we propose a novel multi-document summarization strategy based on Basic Element (BE) vector clustering. In this strategy, sentences are represented by BE vectors instead of word or term vectors before clustering. BE is a head-modifier-relation triple representation of sentence content, and it is more precise to use BE as semantic unit than to use word. The BE-vector clustering is realized by adopting the k-means clustering method, and a novel clustering analysis method is employed to automatically detect the number of clusters, K. The experimental results indicate a superiority of the proposed strategy over the traditional summarization strategy based on word vector clustering. The summaries generated by the proposed strategy achieve a ROUGE-1 score of 0.37291 that is better than those generated by traditional strategy (at 0.36936) on DUC04 task-2.

1 Introduction

With the rapid growth of online information, it becomes more and more important to find and describe textual information effectively. Typical information retrieval (IR) systems have two steps: the first is to find documents based on the user's query, and the second is to rank relevant documents and present them to users based on their relevance to the query. Then the users have to read all of these documents. The problem is that these docs are much relevant and reading them all is time-consuming and unnecessary. Multi-document summarization aims at extracting major information from multiple documents and has become a hot topic in NLP. Multi-document summarization can be classified into three categories according to the way that summaries are created: sentence extraction, sentence compression and information fusion.

The sentence extraction strategy ranks and extracts representative sentences from the multiple documents. Radev [1] described an extractive multi-document summarizer which extracts a summary from multiple documents based on the document cluster centroids. To enhance the coherence of summaries, Hardy Hilda [2]

Deriving Event Relevance from the Ontology Constructed with Formal Concept Analysis

Wei Xu[1,2], Wenjie Li[1], Mingli Wu[1], Wei Li[1], Chunfa Yuan[2]

[1] Department of Computing,
The Hong Kong Polytechnic University, Hong Kong
{cswxu, cswli, csmlwu, cswli}@comp.polyu.edu.hk

[2] Department of Computer Science and Technology
Tsinghua University, China
{vivian00, cfyuan}@mails.tsinghua.edu.cn

Abstract. In this paper, we present a novel approach to derive event relevance from event ontology constructed with Formal Concept Analysis (FCA), a mathematical approach to data analysis and knowledge representation. The ontology is built from a set of relevant documents and according to the named entities associated to the events. Various relevance measures are explored, from binary to scaled, and from symmetrical to asymmetrical associations. We then apply the derived event relevance to the task of multi-document summarization. The experiments on DUC 2004 data set show that the relevant-event-based approaches outperform the independent-event-based approach.

1 Introduction

Extractive summarization is to select the sentences which contain salient concepts in documents. An important issue with it is what criteria should be used to extract the sentences. Event-based summarization attempts to select and organize the sentences in a summary with respect to the events or the sub-events that the sentences describe [1, 2]. As the relevance of events reveals the significance of events, it helps singling out the sentences with the most core events. However, the event-based summarization techniques reported so far explored the events independently.

In the realm of information retrieval, term relations were commonly derived either from a thesaurus like WordNet or from the corpus where the contexts of the terms were investigated. Likewise, mining event relevance requires taking contexts of event happenings into account. The event contexts in our definition are event arguments, such as participants, locations and occurrence times, etc. They are important in defining events and distinguishing them from one another. By this observation, we make use of the named entities associated with the events as event contexts and characterize the events with the verbs and action-denoting nouns prescribed by the named entities.

A sentence compression module for machine-assisted subtitling

Nadjet Bouayad-Agha, Angel Gil,
Oriol Valentin, and Victor Pascual

Universitat Pompeu Fabra,
Barcelona, Spain
nadjet.bouayad@upf.edu, firstname.lastname@upf.edu

Abstract. We present in this paper a sentence compression module used in a machine-assisted subtitling application developed in the European e-content project e-title. Our approach to compression and the architecture of the system are motivated by the commercial and multilingual nature of the project, that is, the need to output reasonable compressions and the ability to add new strategies, genres and languages easily. The compression module currently works for the Catalan and English languages and uses the Constraint Grammar engine for linguistic preprocessing and for the linguistically motivated compression rules, thus providing a homogenous format throughout the compression process. The compression rules were implemented based on a corpus of automatically aligned <script,subtitle> pairs of films for both languages. We performed for both languages an automatic quantitative evaluation of the compression using the aligned corpus and a qualitative manual evaluation of grammaticality and informativeness.

1 Motivation

We present in this paper a sentence compression module used in a machine-assisted subtitling application developed in the European e-content project e-title.[1] This application integrates speech-text synchronisation, machine translation and sentence compression to assist subtitlers in the different stages of the subtitling process. Our approach to compression and the architecture of the system are motivated by the commercial and multilingual nature of the project, that is, the need to output reasonable compressions and the ability to add new strategies, genres and languages easily. We surveyed various approaches to sentence compression developed in Natural Language Processing, for example (Jing, 2000; Zechner, 2001; Hori and Furui, 2002, Knight and Marcu, 2002, Vandeghinste and Pan 2004), and compiled the following list of desiderata for our system:

Grammaticality: the compression module should preserve grammaticality. We found that some approaches guarantee the grammaticality of the output but at the cost of heavier linguistic machinery such as full parsing and subcategorization information (Jing,2000).

[1] EDC22160. Jan.2004–Jan.2006.

Application of Semi-supervised Learning to Evaluative Expression Classification

Yasuhiro Suzuki[1], Hiroya Takamura[2], and Manabu Okumura[2]

[1] Interdisciplinary Graduate School of Science and Engineering,
Tokyo Institute of Technology,
4259 Nagatsuta Midori-ku Yokohama, JAPAN, 226-8503
yasu@lr.pi.titech.ac.jp

[2] Precision and Intelligence Laboratory, Tokyo Institute of Technology,
4259 Nagatsuta Midori-ku Yokohama, JAPAN, 226-8503
{takamura,oku}@pi.titech.ac.jp

Abstract. We propose to use semi-supervised learning methods to classify evaluative expressions, that is, tuples of subjects, their attributes, and evaluative words, that indicate either favorable or unfavorable opinions towards a specific subject. Due to its characteristics, the semi-supervised method that we use can classify evaluative expressions in a corpus by their polarities. This can be accomplished starting from a very small set of seed training examples and using contextual information in the sentences to which the expressions belong. Our experimental results with actual Weblog data show that this bootstrapping approach can improve the accuracy of methods for classifying favorable and unfavorable opinions.

1 Introduction

An increasing amount of work has been devoted to investigating methods of detecting favorable or unfavorable opinions towards specific subjects (e.g., companies and their products) within online documents such as Weblogs (blogs), messages in a chat room and on bulletin board (BBS) [1, 2, 7, 9, 11, 12, 18]. Areas of application for such an analysis are numerous and varied, ranging from analysis of public opinion, customer feedback, and marketing analysis to detection of unfavorable rumors for risk management. The analyses are potentially useful tools for the commercial activities of both companies and individual consumers who want to know the opinions scattered on the World Wide Web (WWW).

To analyze a huge amount of favorable or unfavorable opinions, we need to automatically detect evaluative expressions in text.

Evaluative expressions are not mere words that indicate unique (favorable or unfavorable) polarity in themselves (such as the adjectives 'beautiful' and 'bad'), but rather they are tuples of the subject to be evaluated, an attribute, and an evaluative word. Tuples are necessary because the evaluative polarity of

[†] Yasuhiro Suzuki currently works at Fujitsu.

A New Algorithm for Fast Discovery of Maximal Sequential Patterns in a Document Collection

René A. García-Hernández José Fco. Martínez-Trinidad
Jesús Ariel Carrasco-Ochoa

National Institute of Astrophysics, Optics and Electronics (INAOE)
Puebla, México
{renearnulfo, fmartine, ariel}@inaoep.mx

Abstract. Sequential pattern mining is an important tool for solving many data mining tasks and it has broad applications. However, only few efforts have been made to extract this kind of patterns in a textual database. Due to its broad applications in text mining problems, finding these textual patterns is important because they can be extracted from text independently of the language. Also, they are human readable patterns or descriptors of the text, which do not lose the sequential order of the words in the document. But the problem of discovering sequential patterns in a database of documents presents special characteristics which make it intractable for most of the apriori-like candidate-generation-and-test approaches. Recent studies indicate that the pattern-growth methodology could speed up the sequential pattern mining. In this paper we propose a pattern-growth based algorithm (DIMASP) to discover all the maximal sequential patterns in a document database. Furthermore, DIMASP is incremental and independent of the support threshold. Finally, we compare the performance of DIMASP against GSP, DELISP, GenPrefixSpan and cSPADE algorithms.

1. Introduction

The *Knowledge Discovery in Databases* (KDD) is defined by Fayyad [1] as "the non-trivial process of identifying valid, novel, potentially useful and ultimately understandable patterns in data". The key step in the knowledge discovery process is the data mining step, which following Fayyad: "consisting of applying data analysis and discovery algorithms that, under acceptable computational efficiency limitations, produce a particular enumeration of patterns over the data". This definition has been extended to *Text Mining* like: "consisting of applying text analysis and discovery algorithms that, under acceptable computational efficiency limitations, produce a particular enumeration of patterns over the text". So, text mining is the process that deals with the extraction of patterns from textual data. This definition is used by Feldman [2] to define *Knowledge Discovery in Texts* (KDT). In both KDD and KDT tasks, special attention is required in the performance of the algorithms because they are applied on a large amount of information. In particular the KDT process needs to define simple structures that can be extracted from text documents automatically and in a reasonable time. These structures must be rich enough to allow interesting KD operations [2] having in mind that in some cases the document database is updated.

A Machine Learning Based Approach for Separating Head from Body in Web-Tables[1]

Sung-Won Jung, Hyuk-Chul Kwon

Korean Language Processing Lab. Department of Computer Science and Engineering
Pusan National University, Busan, Korea
{swjung, hckwon}@pusan.ac.kr

Abstract. This study aims to separate the head from the data in web-tables to extract useful information. To achieve this aim, web-tables must be converted into a machine readable form, an attribute-value pair, the relation of which is similar to that of head-body. We have separated meaningful tables and decorative tables in our previous work, because web-tables are used for the purpose of knowledge structuring as well as document design, and only meaningful tables can be used to extract information. In order to extract the semantic relations existing between language contents in a meaningful table, this study separated the head from the body in meaningful tables using machine learning. We (a) established features observing the editing habit of authors and tables themselves, and (b) established a model using machine learning algorithm, C4.5 in order to separate the head from the body. We obtained 86.2% accuracy in extracting the head from the meaningful tables.

1 Introduction

Information extraction encounters various text types. Generally, editors produce three types of text: free text, structured text, and semi-structured text. Among those, free text, composed of natural language sentences, is the most frequently found. To extract information from free text, a computer must analyze the text using natural-language-processing techniques. However, practical application of natural language understanding is still far from being achieved. On the contrary, authors make structured text for specific aims such as a database or a file. These texts contain restricted and well-formed rules. Computers can easily analyze them even though they do not contain structured information apart from that which is predefined. Semi-structured text falls between structured and free text. We can include tables and charts in this type. These texts are easier to analyze and contain more useful and dense information than free text, because of their structural features. This paper focuses on the table among the semi-structured texts, because the table is usually used in HTML documents and easily extracted from HTML documents.

[1] This work was supported by the National Research Laboratory Program M10400000279-05J0000-27910 of Korea Science and Engineering Foundation.

Clustering Abstracts of Scientific Texts using the Transition Point Technique *

David Pinto[1,2], Héctor Jiménez-Salazar[1], and Paolo Rosso[2]

[1] Faculty of Computer Science, BUAP, Puebla 72570,
Ciudad Universitaria, MEXICO
{davideduardopinto, hgimenezs}@gmail.com

[2] Department of Information Systems and Computation,
UPV, Valencia 46022,
Camino de Vera s/n, SPAIN
{dpinto, prosso}@dsic.upv.es

Abstract. Free access to scientific papers in major digital libraries and other web repositories is limited to only their abstracts. Current keyword-based techniques fail on narrow domain-oriented libraries, e.g., those containing only documents on high energy physics like those of the *hep-ex* collection of CERN. We propose a simple procedure to cluster abstracts which consists in applying the transition point technique during the term selection process. This technique uses the mid-frequency terms to index the documents due to the fact that they have a high semantic content. In the experiments we have carried out, the transition point approach has been compared with well known unsupervised term selection techniques. Transition point technique shown that it is possible to obtain a better performance than traditional methods. Moreover, we propose an approach to analyse the stability of transition point term selection method.

1 Introduction

Nowadays, very short text clustering on narrow domains has not received too much attention by the computational linguistic community. This is derived from the high challenge that this problem implies, since the obtained results are very unstable or imprecise when clustering abstracts of scientific papers, technical reports, patents, etc. But, as we can see, most digital libraries and other web-based repositories of scientific and technical information nowadays provide free access only to abstracts and not to the full texts of the documents. Moreover, some institutions, like the well known CERN[1], receive hundreds of publications every day that must be categorized on some specific domain with an unknown

* This work was partially supported by BUAP-VIEP 3/G/ING/05, R2D2 (CICYT TIC2003-07158-C04-03), ICT EU-India (ALA/95/23/2003/077-054), and Generalitat Valenciana Grant (CTESIN/2005/012).

[1] Centre Européen pour la Recherche Nucléaire

Sense Cluster based Categorization and Clustering of Abstracts

Davide Buscaldi[1], Paolo Rosso[1], Mikhail Alexandrov[2], and Alfons Juan Ciscar[1]

[1] Dpto. Sistemas Informáticos y Computación (DSIC),
Universidad Politécnica de Valencia, Spain
{dbuscaldi, prosso, ajuan}@dsic.upv.es
[2] Center for Computing Research,
National Polytechnic Institute, Mexico
dyner1950@mail.ru

Abstract. This paper focuses on the use of sense clusters for classification and clustering of very short texts such as conference abstracts. Common keyword-based techniques are effective for very short documents only when the data pertain to different domains. In the case of conference abstracts, all the documents are from a narrow domain (i.e., share a similar terminology), that increases the difficulty of the task. Sense clusters are extracted from abstracts, exploiting the WordNet relationships existing between words in the same text. Experiments were carried out both for the categorization task, using Bernoulli mixtures for binary data, and the clustering task, by means of Stein's MajorClust method.

1 Introduction

Typical approaches to document clustering and categorization in a given domain are to transform the textual documents into vector form, by using a list of index keywords. This kind of approaches has also been used for clustering etherogeneous short documents (e.g. documents containing 50-100 words) with good results. However, term-based approaches usually give unstable or imprecise results when applied to documents from one narrow domain.

Previous works on narrow-domain short document classification obtained good results by using supervised methods and set of keywords (*itemsets*) as index terms [3].

In this work, we exploited the linguistic information extracted from WordNet in order to extract key *concept clusters* from the documents, using the method proposed by Bo-Yeong Kang *et al.* [5], which is based on semantic relationships between the terms in the document. Concept clusters are used as index words.

Various methods have been tested for the categorization and clustering task, including Bernoulli mixture models, which have been investigated for text categorization in [4]. Text categorization procedures are based on either binary or integer-valued features. In our case, due to the low absolute frequency observable in short documents, we used only the information if an index term was or not in the abstract, thus obtaining a binary representation of each document.

Analysing part-of-speech for Portuguese Text Classification

Teresa Gonçalves[1], Cassiana Silva[2], Paulo Quaresma[1], and Renata Vieira[2]

[1] Dep. Informática, Universidade de Évora, 7000 Évora, Portugal
tcg,pq@di.uevora.pt
[2] Unisinos, CEP 93.022-000 São Leopoldo, RS, Brasil
cassiana,renata@exatas.unisinos.br

Abstract. This paper proposes and evaluates the use of linguistic information in the pre-processing phase of text classification. We present several experiments evaluating the selection of terms based on different measures and linguistic knowledge. To build the classifier we used Support Vector Machines (SVM), which are known to produce good results on text classification tasks.
Our proposals were applied to two different datasets written in the Portuguese language: articles from a Brazilian newspaper (Folha de São Paulo) and juridical documents from the Portuguese Attorney General's Office. The results show the relevance of part-of-speech information for the pre-processing phase of text classification allowing for a strong reduction of the number of features needed in the text classification.

1 Introduction

Machine learning techniques are applied to document collections aiming at extracting patterns that may be useful to organise or retrieve information from large collections. Tasks related to this area are text classification, clustering, summarisation, and information extraction. One of the first steps in text mining tasks is the pre-processing of the documents, as they need to be represented in a more structured way to be fed to machine learning algorithms. In this step, words are extracted from the documents and, usually, a subset of words (stop words) is not considered, because their role is related to the structural organisation of the sentences and does not have discriminating power over different classes. This shallow and practical approach is known as bag-of-words. Usually, to reduce semantically related terms to the same root, a lemmatiser is applied.

Finding more elaborated models is still a great research challenge in the field; natural language processing increases the complexity of the problem and these tasks, to be useful, require efficient systems. Our proposal considers that there is still lack of knowledge about how to bring natural language and traditionally known techniques of data mining tasks together for efficient text mining. Therefore, here we make an analysis of different word categories (nouns, adjectives, proper names, verbs) for text mining, and perform a set of experiments of

Improving kNN Text Categorization by Removing Outliers from Training Set [*]

Kwangcheol Shin, Ajith Abraham, and Sang Yong Han[+]

School of Computer Science and Engineering, Chung-Ang University
221, Heukseok-dong, Dongjak-gu, Seoul 156-756, Korea
kcshin@archi.cse.cau.ac.kr,
ajith.abraham@ieee.org, hansy@cau.ac.kr

Abstract. We show that excluding outliers from the training data significantly improves kNN classifier, which in this case performs about 10% better than the best know method—Centroid-based classifier. Outliers are the elements whose similarity to the centroid of the corresponding category is below a threshold.

1 Introduction

Since late 1990s, the explosive growth of Internet resulted in a huge quantity of documents available on-line. Technologies for efficient management of these documents are being developed continually. One of representative tasks for efficient document management is text categorization, called also classification: given a set of training examples assigned each one to some categories, to assign new documents to a suitable category.

A well-known text categorization method is kNN [1]; other popular methods are Naive Bayesian [3], C4.5 [4], and SVM [5]. Han and Karypis [2] proposed the Centroid-based classifier and showed that it gives better results than other known methods.

In this paper we show that removing outliers from the training categories significantly improves the classification results obtained with kNN method. Our experiments show that the new method gives better results than the Centroid-based classifier.

2 Related Work

Document representation. In both categorization techniques considered below, documents are represented as keyword vectors according to the standard vector space model with *tf-idf* term weighting [6, 7]. Namely, let the document collection contains

[*] Work supported by the MIC (Ministry of Information and Communication), Korea, under the Chung-Ang University HNRC-ITRC (Home Network Research Center) support program supervised by the IITA (Institute of Information Technology Assessment).

[+] Corresponding author.

Writing for Language-Impaired Readers

Aurélien Max

LIMSI-CNRS & Université Paris Sud
Bâtiment 508, F-91405 Orsay Cedex, France
aurelien.max@limsi.fr

Abstract. This paper advocates an approach whereby the needs of language-impaired readers are taken into account at the stage of text authoring by means of NLP integration. In our proposed system architecture, a simplification module produces candidate simplified rephrasings that the author of the text can accept or edit. This article describes the syntactic simplification module which has been partly implemented. We believe the proposed approach constitutes a framework for the more general task of authoring NLP-enriched documents obtained through validations from the author.

1 Introduction

The goal of NLP, as it seems, is mainly to do some processing on existing text. Another domain of application that we believe has a great potential is the use of NLP during text creation, where it can help authors write better documents in a more efficient way. A lot of the difficulties when processing real-life documents arise from the inherent complexity of natural language, which requires word-sense and syntactic structure disambiguation, to name just two. In fact, rule-based and statistical NLP systems are rather good at finding hypotheses, but they often fail when it comes to ranking them and finding the appropriate solution in context.

Some cases can certainly justify the extra cost of annotating the text with the result of the correct analysis, thus permitting much better results on NLP tasks. This concept has already been investigated, for example in the Dialogue-based Machine Translation paradigm [1] whereby a monolingual writer answers ambiguity questions. This process yields a disambiguated analysis for each sentence that is then sent to a machine translation engine.

The kinds of annotation that can be obtained through interaction can be of very different natures. One kind is a transformation of the initial text: for example, annotations at the paragraph level can be assembled to constitute a summary. Transformations at the sentence level include paraphrasing and its differents uses. Among them, text simplification has attracted significant interest in the past years [4, 3, 5]. The most notable application of text simplification has been as an assistive technology for people suffering from aphasia, a loss of language that can result in severe comprehension disorders. It is very unlikely that a text transformation system could produce a coherent text conveying the

Document Copy Detection System based on Plagiarism Patterns*

NamOh Kang, SangYong Han†

School of Computer Science and Engineering
ChungAng University, Seoul, Korea
kang@archi.cse.cau.ac.kr, hansy@cau.ac.kr

Abstract. Document copy detection is a very important tool for protecting author's copyright. We present a document copy detection system that calculates the similarity between documents based on plagiarism patterns. Experiments were performed using CISI document collection and show that the proposed system produces more precise results than existing systems.

1 Introduction

For protecting author's copyright, many kinds of intellectual property protection techniques have been introduced; copy prevention, signature and content based copy detection, etc. Copy protection and signature-based copy detection can be very useful to prevent or detect copying of a whole document. However, these techniques have some drawbacks that they make it difficult for users to share information and can not prevent copying of the document in partial parts [1].

Huge amount of digital documents is made public day to day in Internet. Most of the documents are not supported by either copy prevention technique or signature based copy detection technique. This situation increases the necessity in content based copy detection techniques. So far, many document copy detection (DCD) systems based on content based copy detection technique have been introduced, for example COPS [2], SCAM [1], CHECK [3], etc. However, most DCD systems mainly focus on checking the possibility of copy between original documents and a query document. They do not give any evidence of plagiaristic sources to user. In this paper, we propose a DCD system that provides evidence of plagiarism style to the user.

2 Comparing Unit and Overlap Measure Function

DCD system divides documents efficiently in comparing unit (chunking unit) for checking the possibility of copy. In this paper, we select the comparing unit as a

* This research was supported by the MIC (Ministry of Information and Communication), Korea, under the Chung-Ang University HNRC-ITRC (Home Network Research Center) support program supervised by the IITA (Institute of Information Technology Assessment).

† Corresponding author.

Regional vs. global robust spelling correction

M. Vilares[1], J. Otero[1], and V.M. Darriba[1]

Department of Computer Science, University of Vigo
Campus As Lagoas s/n, 32004 Orense, Spain
{vilares,jop,darriba}@uvigo.es

Abstract. We explore the practical viability of a regional architecture to deal with robust spelling correction, a process including both unknown sequences recognition and spelling correction. Our goal is to reconcile these techniques from both the topological and the operational point of view. In contrast to the global strategy of most spelling correction algorithms, and local ones associated with the completion of unknown sequences, our proposal seems to provide an unified framework allowing us to maintain the advantages in each case, and avoid the drawbacks.

1 Introduction

In describing human performance in spelling correction, as compared to machine performance, we should try to take into account both the computational efficiency and the quality achieved in order to equal, or even do better than humans. This translates into a trade-off between the study of the often complex linguistic phenomena involved and the efficiency of the operational mechanisms available for implementation. In order to attain this goal, simple proposals can be sufficient to overcome most limits to providing an efficient strategy, even in the case of interactive applications. In fact, most approaches are oriented to improving first-guess accuracy [1] and/or to considering filter-based solutions to speed up the process [8]. So, system developers expect to reduce the time needed to ensure an adequate coverage of the problem, before taking into account more sophisticated linguistic information.

The state of the art techniques mainly focus on approximate string matching proposals, often firstly oriented to searching [2], although they can be easily adapted to robust spelling correction tasks [4]. Essentially, these algorithms apply a metric [5] to measure the minimun number of unit operations are necessary to convert one string into another, sometimes embedding this task in the recognizer [10] in order to improve the computational efficiency. In this context, we identify a set of objective parameters in order to evaluate different approaches and algorithms in dealing with robust spelling correction.

[*] This research has been partially supported by the Spanish Government under projects TIN2004-07246-C03-01, and the Autonomous Government of Galicia under projects PGIDIT05PXIC30501PN, PGIDIT05SIN059E, PGIDIT05SIN044E, PGIDIT04SIN065E and PGIDIT03SIN30501PR.

Author Index