Advances in Natural Language Processing

Research in Computing Science

Research in Computing Science es una publicación trimestral, de circulación internacional, editada por el Centro de Investigación en Computación del IPN, para dar a conocer los avances de investigación científica y desarrollo tecnológico de la comunidad científica internacional. **Volumen 18**, Febrero, 2006. Tiraje: 500 ejemplares. *Certificado de Reserva de Derechos al Uso Exclusivo del Título* No. 04-2004-062613250000-102, expedido por el Instituto Nacional de Derecho de Autor. *Certificado de Licitud de Título* No. 12897, *Certificado de licitud de Contenido* No. 10470, expedidos por la Comisión Calificadora de Publicaciones y Revistas Ilustradas. El contenido de los artículos es responsabilidad exclusiva de sus respectivos autores. Queda prohibida la reproducción total o parcial, por cualquier medio, sin el permiso expreso del editor, excepto para uso personal o de estudio haciendo cita explícita en la primera página de cada documento. Imagen de la portada: www.absolutewallpapers.com/wallpapers/3dwallpapers/fractal/fractal_8.jpg. Impreso en la Ciudad de México, en los Talleres Gráficos del IPN – Dirección de Publicaciones, Tres Guerras 27, Centro Histórico, México, D.F. Distribuida por el Centro de Investigación en Computación, Av. Juan de Dios Bátiz S/N, Esq. Av. Miguel Othón de Mendizábal, Col. Nueva Industrial Vallejo, C.P. 07738, México, D.F. Tel. 57 29 60 00, ext. 56571.

Editor Responsable: *Juan Humberto Sossa Azuela,* RFC *SOAJ560723*

Research in Computing Science *is published by the Center for Computing Research of IPN.* ***Volume 18****, February, 2006. Printing 500. The authors are responsible for the contents of their articles. All rights reserved. No part of this publication may be reproduced, stored in a retrieval system, or transmitted, in any form or by any means, electronic, mechanical, photocopying, recording or otherwise, without prior permission of Centre for Computing Research. Printed in Mexico City, February, 2006, in the IPN Graphic Workshop – Publication Office.*

Volume 18

Volumen 18

Advances in Natural Language Processing

Volume Editor:
Editor del Volumen

Alexander Gelbukh

Instituto Politécnico Nacional
Centro de Investigación en Computación
México 2006

ISSN: 1665-9899

Instituto Politécnico Nacional (IPN)
Centro de Investigación en Computación (CIC)
Av. Juan de Dios Bátiz s/n esq. M. Othón de Mendizábal
Unidad Profesional "Adolfo López Mateos", Zacatenco
07738, México D.F., México

http://www.ipn.mx
http://www.cic.ipn.mx

Printing: 500
Impresiones: 500

Printed in Mexico
Impreso en México

Preface

Natural Language Processing is a branch of Artificial Intelligence aimed at enabling the computers to perform meaningful human-like activities related to written or spoken human language, such as English or Spanish. For traditional computer programs, a text is just a long string of characters—which the program can copy, print, or erase. In contrast, intelligent natural language processing programs are designed for operations involving the meaning of the text.

The most important applications of natural language processing include information retrieval and information organization, machine translation, and natural language interfaces, among others. However, as in any science, the activities of the researchers are mostly concentrated on its internal art and craft, that is, on the solution of the problems arising in analysis or synthesis of natural language text or speech, such as syntactic and semantic analysis, disambiguation, or compilation of dictionaries and grammars necessary for such analysis.

This volume presents 19 original research papers written by 63 authors representing 16 different countries: Argentina, Austria, Canada, China, Finland, Hong Kong, India, Japan, Korea, Lebanon, Mexico, Norway, Spain, UK, United Arab Emirates, and USA. The volume is structured in 7 thematic areas of both theory and applications of Natural Language Processing:

- Lexical Resources
- Morphology and syntax
- Word Sense Disambiguation and Semantics
- Speech Processing
- Information Retrieval
- Question Answering and Text Summarization
- Computer-Assisted Language Learning

The papers included in this volume were selected on the base of rigorous international reviewing process out of 43 submissions received for evaluation; thus the acceptance rate of this volume was 44%.

I would like to cordially thank all people involved in the preparation of this volume. In the first place I want to thank the authors of the published paper for their excellent research work giving sense to the work of all other people involved, as well the authors of rejected papers for their interest and effort. I also thank the members of the Editorial Board of the volume and additional reviewers for their hard work on reviewing and selecting the papers. I thank Sulema Torres, Ignacio Garcia Araoz, Oralia del Carmen Pérez Orozco, and Raquel López Alamilla for their valuable collaboration in preparation of this volume. The submission, reviewing, and selection process was supported for free by the EasyChair system, www.EasyChair.org.

Alexander Gelbukh February 2006

Table of Contents

Índice

Page/Pág.

Lexical Resources

Morphology and syntax

Word Sense Disambiguation and Semantics

Speech Processing

Information Retrieval

Question Answering and Text Summarization

Computer-Assisted Language Learning

Generating Multilingual Grammars from OWL Ontologies

Guillermo Pérez, Gabriel Amores, Pilar Manchón and David González

Universidad de Sevilla
Seville, Spain
{gperez, jgabriel, pmanchon, dgmaline}@us.es

Abstract. This paper describes a tool which automatically generates productions for context–free grammars from OWL ontologies, using just a rule–based configuration file. This tool has been implemented within the framework of a dialogue system, and has achieved several goals: it leverages the manual work of the linguist, and ensures coherence and completeness between the Domain Knowledge (Knowledge Manager Module) and the Linguistic Knowledge (Natural Language Understanding Module) in the application.

1 Introduction

1.1 Automatic Grammar Generation

The problem of manually generating grammars for a Natural Language Understanding (NLU) system has been widely discussed. Two main approaches can be highlighted from those proposed in the literature: Grammatical Inference and Rule Based Grammar Generation.

The Grammatical Inference approach (http://eurise.univ-st-etienne.fr/gi/) refers to the process of learning grammars and languages from data and is considered nowadays as an independent research area within Machine Learning techniques. Examples of applications based on this approach are ABL [1] and EMILE [2].

On the other hand, the Rule Based approach tries to generate the grammar rules from scratch (i.e. based on the expertise of a linguist), while trying to minimize the manual work. An example of this approach is the Grammatical Framework [3], whose proposal is to organize the multilingual grammar construction in two building blocks: an abstract syntax which contains category and function declarations, and a concrete syntax which defines linearization rules. Category and function declarations are shared by all languages and thus appear only once, while linearization rules are defined on a per–language basis. Methods which generate grammars from ontologies (including ours) may also be considered examples of the Rule Based approach.

© A. Gelbukh (Editor)
Advances in Natural Language Processing
Research in Computing Science 18, 2006, pp. 3-14

Clustering of English-Korean Translation Word Pairs Using Bi-grams

Hanmin Jung[1], Hee-Kwan Koo[2], Won-Kyung Sung[1] and Dong-In Park[1]

[1] NTIS Division, KISTI, Korea
jhm@kisti.re.kr
[2] Practical Information Science, UST, Korea

Abstract. This paper describes a clustering algorithm for Korean translation words automatically extracted from Korean newspapers. Since above 80% of English words appear with abbreviated forms in Korean newspapers, it is necessary to make the clusters of their Korean translation words to easily construct bi-lingual knowledge bases such as dictionaries and translation patterns. As a seed to acquire each translation cluster, we repeat to choose an adequate translation word from a remaining translation set using an extended bi-gram-based binary vector matching until the set becomes empty. We also deal with several phenomena such as transliterations and acronyms during the clustering. Experimental results showed that our algorithm is superior to Dice coefficient and Jaccard coefficient in both determining adequate translation words and clustering translations.

1 Introduction

As information technology develops in recent years, many terminologies are rapidly created and discarded. Newspapers are excellent resources to acquire new-coined terms and to inspect their life cycle [4]. About 90% of terms in Korean newspapers, in particular, are originated from foreign languages such as English and Chinese[1] [1]. Some of them are accompanied by original words in English for readers to easily grasp the meaning, for example, "세계무역기구 (WTO)." However, many English words (about 82% in our test set) appear with abbreviated forms, and translations differ like "아시아태평양경제협력기구," "아시아태평양경제협력체," "아태경제협력체," and "아태경제협력회의" for "APEC; Asia-Pacific Economic Cooperation." Such English abbreviated forms tend to cause word sense ambiguities, for example, "Internet Service Provider," "Information Strategic Planning," and "Image Signal Processor" for "ISP." Newspapers also usually use parentheses to represent a pair of translation pairs, but they are not limited to the pairs. Many extraction errors are caused by the free uses of parentheses such as "모델명 S3C2410 (CPU)"[2] and

[1] E.g., "아펙" is a Korean transliterated word for English "APEC," and "경제" is for Chinese "經濟."
[2] "모델명 S3C2410" = "모델명 (Model No.)" + "S3C2410."

Advances in Natural Language Processing
Research in Computing Science 18, 2006, pp. 15-24

Modified Makagonov's Method for Testing Word Similarity and its Application to Constructing Word Frequency Lists

Xavier Blanco,[1] Mikhail Alexandrov,[1,2] and Alexander Gelbukh [2]

[1] Department of French and Romance Philology, Autonomous University of Barcelona
dyner1950@mail.ru, Xavier.Blanco@uab.es

[2] Center for Computing Research, National Polytechnic Institute (IPN), Mexico
dyner@cic.ipn.mx; www.Gelbukh.com

Abstract. By (morphologically) similar wordforms we understand wordforms (strings) that have the same base meaning (roughly, the same root), such as *sadly* and *sadden*. The task of deciding whether two given strings are similar (in this sense) has numerous applications in text processing, e.g., in information retrieval, for which usually stemming is employed as an intermediate step. Makagonov has suggested a weakly supervised approach for testing word similarity, based on empirical formulae comparing the number of equal and different letters in the two strings. This method gives good results on English, Russian, and a number of Romance languages. However, his approach does not deal well with slight morphological alterations in the stem, such as Spanish *pensar* vs. *pienso*. We propose a simple modification of the method using n-grams instead of letters. We also consider four algorithms for compiling a word frequency list relying on these formulae. Examples from Spanish and English are presented.

1 Introduction

Given a large text or text corpus and a pair of wordforms (strings), we consider the task of guessing whether these two words have the same root and thus the same base meaning. We call this (morphological) *word similarity*: two words are *similar* if they have the same root. This relation permits grouping together the words having the same root, e.g., *sad*, *sadly*, *sadness*, *sadden*, *saddened*, etc. This task has numerous applications, such as constructing word frequency lists. Our motivation is to improve information retrieval and similar practical applications. Consequently, our goal is to provide a reasonably accurate statistical-based algorithm (tolerating certain error rate) and not a precise linguistic analysis.

For grouping together the words with the same root, two morphology-based methods are usually used: lemmatization and stemming. Lemmatization reduces words to the base form: *having* → *have*; stemming truncates words to their stems: *having* → *hav-* (often lemmatization task is also referred to as stemming).

Stemming or lemmatization can be used for testing the (morphological) similarity between two words: both words are first reduced to lemmas or stems; if the resulting strings are equal then the two given words are declared similar. This gives a symmet-

Advances in Natural Language Processing
Research in Computing Science 18, 2006, pp. 27-36

HMM based POS Tagger for a Relatively Free Word Order Language

Arulmozhi Palanisamy and Sobha Lalitha Devi

AU-KBC Research Centre, MIT Campus,
Chromepet, Chennai - 600044, India.
{p.arulmozhi,sobha}@au-kbc.org

Abstract. We present an implementation of a part-of-speech tagger based on hidden markov model for Tamil, a relatively free word order, morphologically productive and agglutinative language. In HMM we assume that probability of an item in a sequence depends on its immediate predecessor. That is the tag for the current word depends up on the previous word and its tag. Here in the state sequence the tags are considered as states and the transition from one state to another state has a transition probability. The emission probability is the probability of observing a symbol in a particular state. In achieving this, we use viterbi algorithm. The basic tag set including the inflection is 53. Tamil being an agglutinative language, each word has different combinations of tags. Compound words are also used very often. So, the tagset increases to 350, as the combinations become high. The training corpus is tagged with the combination of basic tags and tags for inflection of the word. The evaluation gives encouraging result.

1 Introduction

Part of speech (POS) tagging is the task of labeling each word in a sentence with its appropriate syntactic category called part of speech. It is an essential task for all the language processing activities. There are two factors determining the syntactic category of a word (a) the words lexical probability (e.g. without context, *bank* is more probably a noun than a verb) (b) the words contextual probability (e.g. after a noun or a pronoun, bank is more a verb than a noun, as in "I bank with HSBC "). Hence it disambiguates the part of speech of a word when it occurs in different contexts. For any POS work the tagset of the language has to be developed. It should contain the major tags and the morpho-syntactic features called the sub tags.

In this paper we present a POS tagger developed for Tamil using HMM and Viterbi transition. Tamil is a South Dravidian language, which is relatively free word order, morphologically productive and agglutinative in nature. It is an old language and most of the words are built up from roots by following certain patterns and adding suffixes. Due to the agglutination many combination of words and suffixes occur to form a single word whose POS is a combination of all the suffixed words or suffixes. Hence the total number of tagset increases as the combination increases. In this work we confine to a fixed set of tagset, which gives the most commonly needed tags for any

Advances in Natural Language Processing
Research in Computing Science 18, 2006, pp. 37-48

Extended Tagging in Requirements Engineering

Günther Fliedl, Christian Kop, Heinrich C. Mayr,
Jürgen Vöhringer, Georg Weber, Christian Winkler

University of Klagenfurt

Abstract. In this paper standard tagging mechanisms are discussed and partly criticized. We propose a semantically and morphosyntactically enriched mechanism in which many of the shortcomings of standard POS-taggers are eliminated. Therefore rule based disambiguation steps are presented. They include mainly a specification of contextually motivated verbal subcategories. We need this fine grained system for better preprocessing of requirements texts which have to be highly explicit and non ambiguous. This linguistic preprocessor can support an interpretation tool used to extract semantic concepts and relations for the requirements engineering step in software development.

1 Introduction

Classical tagging approaches use standardized (POS) tag sets. Such kind of standardized tagging (e.g., Brilltagger [1], TnT [2], Q-Tag [11] Treetagger [8], [10] etc.), however, show weakness in the following three aspects:

- Tags like 'VAINF' provide only basic categorial and morphological information;
- Ambiguity cannot be made explicit;
- Chunking and identification of multiple tokens is not possible.

To avoid such deficiencies, we developed a system called NIBA[1]-Tag which allows tagging of German with an extended tagset, and inheritance of morphosyntactic and morphosemantic features. Morphosemantic tagging in our sense is labeling words by morpho-syntactically relevant semantic classifiers (sem-tags like 'tvag2', 'eV', 'indef', 'poss', etc.; see Appendix 1 for a rough comparison with the Treebank [7] STTS [S*99]), It has proved to be an efficient method for extracting different types of linguistically motivated information coded in text. The XML-Tagger-output for the German PP (=P2) ***Bei Eintreffen des Auftrags*** in Table 1 shows how the tagging result is structured.

As can be seen in Table 1, our tagging system has the following linguistic competence:

[1] NIBA is a German acronym for Natural language based Requirements Engineering. The project NIBA is supported by the "Klaus Tschira Stiftung Heidelberg".

Advances in Natural Language Processing
Research in Computing Science 18, 2006, pp. 49-56

Analogical Modeling with Bias — allowing Feedback and Centering

Christer Johansson and Lars G. Johnsen

Dept. of Linguistics and Literature
University of Bergen
N-5007 Bergen, Norway
{christer.johansson, lars.johnsen}@lili.uib.no

Abstract. We show a computationally efficient approximation (cf. [1]) of a full analogy model [2, 3], implemented in a computer program, and tested on the CoNLL2000 chunk tagging task [4], putting clause boundaries around mainly np and vp phrases. Our implementation showed to be competitive with other memory based learners. It deviates only slightly from the theoretical model. First, it implements a version of homogeneity check, which does not account fully for nondeterministic homogeneity. Second, it allows feedback of the last classification, and thirdly it allows centering on some central feature positions. Positions containing a) those parts-of-speech tags and b) those words that are to be given a chunk tag are given a weight which is given by how many match patterns that are equally or more general. A match on two centered features gives its patterns an extra weight given by the number of features. The results can be summarized as follows: a) using only lexical features performs below baseline. b) The implementation without anything extra, performs as the baseline for five parts-of-speech features, and centering improves the results. c) Feedback on its own does not improve results, while feedback + centering improves results more than just centering. Feedback on its own makes results deteriorate. The results exceed F=92, which is comparable with some of the best reported results for Memory Based Learning on the chunk tagging task.

1 Introduction

Analogical modeling (AM) is a (memory based) method to evaluate the analogical support for a classification [2, 3, 5]. Chandler [6] suggested AM as an alternative to both rule based and connectionist models of language processing and acquisition. AM defines a natural statistic, which can be implemented by comparisons of subsets of linguistic variables, without numerical calculations [5]. The natural statistic works as a selection mechanism, selecting those patterns in the database which most clearly points out a class for a novel pattern.

The original AM model compares all subsets of investigated variables. This may cause an exponential explosion in the number of comparisons, which has made it difficult to investigate large models with many variables (> 10) combined

© A. Gelbukh (Editor)
Advances in Natural Language Processing
Research in Computing Science 18, 2006, pp. 57-68

Word Sense Disambiguation with Basic-Level Categories

Steve Legrand

Department of Computer Science, University of Jyväskylä, Finland
stelegra@cc.jyu.fi

Abstract. Research in basic-level categories has provided insights that can be beneficially used in word-sense disambiguation. Human beings favor certain categories over others and this is reflected in their use of language. Parts and functions categories are especially useful in providing contextual clues. The disambiguation effort in this article concentrates on the two main senses of the word "palm." The results are encouraging and indicate that basic-level categories will have a role to play in computational linguistics.

1 Introduction

If a word has only one sense, a non-native speaker can confirm its meaning by a quick look at a dictionary. Most of the words do have, however, more than one sense, and both the native and the non-native speaker need to use the word context in order to find its correct sense. For example, when we look at the sentence,

There was a large blister on the heel of his right palm.

it is obvious to us that the word *palm* refers to a body part rather than to a tree or a handheld computer. The words *blister*, *heel*, *his*, and *right* when combined in a certain way point us towards the correct meaning.

Most of the automated disambiguation techniques, one way or another, are context-based, making use not only of the words themselves, but also of the part-of-speech information, word order, document genre and so on. Generally, we can say that these techniques are justified by our observations that certain words do co-occur quite regularly with each other within certain contexts. This notion has been used somewhat heuristically in automated word sense disambiguation, and often there is no reference to any cognitive disambiguation mechanism that could have been involved. Nevertheless, it is not disputed that context plays a very important part in the word sense disambiguation by our cognitive faculties.

The question arises: what is this human disambiguation mechanism like if it exists, and would it be possible to mimic and exploit it in automated word sense disambiguation? Is it rooted in our biology, and consequently reflected in our cognitive abilities, including our ability to categorize? The classical view of categories is often interpreted as meaning that things belong to the same category only if they have certain properties in common. It might seem that car parts such as a wheel and an engine do not share any properties, therefore should one assume that they cannot belong to the

© *A. Gelbukh (Editor)*
Advances in Natural Language Processing
Research in Computing Science 18, 2006, pp. 71-82

A new proposal of Word Sense Disambiguation for nouns on a Question Answering System*

S. Ferrández[1], S. Roger[1,2], A. Ferrández[1], A. Aguilar[1] and P. López-Moreno[1]

[1] Natural Language Processing and Information Systems Group
Department of Software and Computing Systems
University of Alicante, Spain
[2] Natural Language Processing Group
Department of Computing Sciences
University of Comahue, Argentina
{sferrandez,sroger,antonio}@dlsi.ua.es,
{P.Lopez}@ua.es

Abstract. This paper describes the impact of the application of a Word Sense Disambiguation (WSD) algorithm for nouns on AliQAn [16], the Question Answering system with which we have participated in the CLEF-2005. Applying the traditional WSD decreases the performance in 4.7% on the Mean Reciprocal Rank (MRR). To solve this problem, we propose two different uses of WSD: (1) to choose a set of synsets instead of the traditional use of WSD, in which only one synset is chosen; (2) to disambiguate the words not present in EuroWordNet (EWN). Using our proposal of WSD the MRR increases a 6.3% with regard to the baseline without WSD. Furthermore, our proposal of WSD increases the MRR with regard to the traditional use of WSD in an 11%. Finally, the implementation of our approach of WSD is computationally efficient by means of a preprocessing of EWN.

1 Introduction

In this paper we analyze the benefits of a Word Sense Disambiguation (WSD) algorithm for nouns in AliQAn [16], a Spanish Question Answering (QA) system, with which we have participated in the CLEF-2005[1] competition.

QA objective consists of identifying the answer of questions in large collections of documents. QA is not a simple task of Information Retrieval (IR), QA tries to go beyond and returns a concrete answer in response to an arbitrary query. For the users, it is very interesting to find accurate information, thanks to the increment of available information. The QA systems are capable to answer questions formulated by the users in natural language.

* This research has been partially funded by the Spanish Government under project CICyT number TIC2003-07158-C04-01 and by the Valencia Government under project number GV04B-268.

[1] http://www.clef-campaign.org/

© A. Gelbukh (Editor)
Advances in Natural Language Processing
Research in Computing Science 18, 2006, pp. 83-92

On Types, Intension and Compositionality

Walid S. Saba

Department of Computer Science
American University of Technology (AUT)
Byblos-Highway, Byblos, Lebanon P. O. Box 20
walid.saba@aut.edu

Abstract. In this paper we demonstrate that a number of challenging problems in the semantics of natural language, namely the treatment of the so-called intensional verbs and the semantics of nominal compounds, can be adequately resolved in the framework of compositional semantics, if a strongly-typed ontological structure is assumed. In addition to suggesting a proper treatment of nominal compounds and intensional verbs within the framework of compositional semantics, we briefly discuss the nature of this ontological type system and how it may be constructed.

1 The Semantics of Nominal Compounds

The semantics of nominal compounds have received considerable attention by a number of authors, most notably (Kamp & Partee, 1995; Fodor & Lepore, 1996; Pustejovsky, 2001), and to our knowledge, the question of what is an appropriate semantics for nominal compounds has not yet been settled. In fact, it seems that the problem of nominal compounds has presented a major challenge to the general program of compositional semantics in the Montague (1973) tradition, where the meaning of a compound nominal such as $[N_1\ N_2]$ is generally given as follows:

(1) $$\|N_1\ N_2\| = F\left(\|N_1\|, \|N_2\|\right)$$

In the simplest of cases, the compositional function F is usually taken to be a conjunction (or intersection) of predicates (or sets). For example, assuming that $red(x)$ and $apple(x)$ represent the meanings of *red* and *apple*, respectively, then the meaning of a nominal such as *red apple* is usually given as

(2) $$\|\text{red apple}\| = \left\{x \,|\, red(x) \wedge apple(x)\right\}$$

What (2) says is that something is a *red apple* if it is *red* and *apple*. This simplistic model, while seems adequate in this case (and indeed in many other instances of

© *A. Gelbukh (Editor)*
Advances in Natural Language Processing
Research in Computing Science 18, 2006, pp. 93-102

Acoustic Model Adaptation for Codec Speech based on Learning-by-Doing Concept

Shingo Kuroiwa, Satoru Tsuge, Koji Tanaka, Kazuma Hara and Fuji Ren

Faculty of Engineering, The University of Tokushima,
Tokushimashi 770-8506, Japan
{kuroiwa,tsuge,ren}@is.tokushima-u.ac.jp,
WWW home page: http://a1-www.is.tokushima-u.ac.jp/

Abstract. Recently, personal digital assistants like cellular phones are shifting to IP terminals. The encoding-decoding process utilized for transmitting over IP networks deteriorates the quality of speech data. This deterioration causes degradation in speech recognition performance. Acoustic model adaptations can improve recognition performance. However, the conventional adaptation methods usually require a large amount of adaptation data. In this paper, we propse a novel acoustic model adaptation technique that generates "speaker-independent" HMM for the target environment based on the learning-by-doing concept. The proposed method uses HMM-based speech synthesis to generate adaptation data from the acoustic model of HMM-based speech recognizer, and consequently does not require any speech data for adaptation. By using the generated data after coding, the acoustic model is adapted to codec speech. Experimental results on G.723.1 codec speech recognition show that the proposed method improves speech recognition performance. A relative word error rate reduction of approximately 12% was observed.
Keywords: Speech Recognition, Model Adaptation, Codec Speech, Speech Synthesis, Learning-by-Doing

1 Introduction

In recent years, telephone speech recognition systems encompassing thousands of vocabularies have become practical and widely used [1, 2]. These systems are generally utilized by automatic telephone services for booking an airline ticket, inquiring about stock, receiving traffic information, and so on. However, the recognition accuracy of cellular phones is still inadequate due to compression coding or ambient noise [3–5].

Recently, personal digital assistants like cellular phones are shifting to IP terminals. For transmission over IP networks, speech data must be encoded at the sending end and subsequently decoded at the receiving end. This coding process deteriorates the quality of the voice data. Although most people can not notice this deterioration, it seriously affects the performance of those speech recognizers not designed for low-quality voice data[6]. The major causes of speech recognition performance degradation are : distortion in the transmission environment (transmission error and packet loss), and low bitrate speech coding (loss

Advances in Natural Language Processing
Research in Computing Science 18, 2006, pp. 105-114

Specific Speaker's Japanese Speech Corpus over Long and Short Time Periods

Satoru Tsuge[1], Masami Shishibori[1], Fuji Ren[1],
Kenji Kita[2] and Shingo Kuroiwa[1]

[1] Faculty of Engineering, the University of Tokushima,
2-1, Minami Josanjima, Tokushima, Japan
tsuge@is.tokushima-u.ac.jp
[2] Center for Advanced Information Technology, the University of Tokushima,
2-1, Minami Josanjima, Tokushima, Japan

Abstract. It is known that speech recognition performance varies pending when the utterance was uttered although speakers use a speaker-dependent speech recognition system. This implies that the speech varies even if a specific speaker utters a specific sentence. Hence, we investigate the speech variability of a specific speaker over short and long time periods for getting the stable speech recognition performances. For this investigation, we need a specific speaker's speech corpus which is recorded over long time periods. However, at present, we have not seen such a Japanese speech corpus. Therefore, we have been collecting the Japanese speech corpus for investigating the relationship between intra-speaker speech variability and speech recognition performance. In this paper, first, we introduce our speech corpus. Our corpus consists of six speakers' speech data. Each speaker read specific utterance sets tree times a day, once a week. Using a specific female speaker's speech data in this corpus, we conduct speech recognition experiments for investigating the relationship between intra-speaker speech variability and speech recognition performance. Experimental results show that the variability of recognition performance over different days is larger than variability of recognition performance within a day.

1 Introduction

Recently, speech recognition systems, such as car navigation systems, and cellular phone systems have come into wide use. Although a speaker uses a speaker-dependent speech recognition system, it is known that speech recognition performance varies pending when the utterance was uttered. For this reason, we consider that speech characteristics varies even though the speaker and utterance remain constant. This intra-speaker variability is caused by some factors including emotion and background noise. If the recognition performance is not consistent, then products using speech recognition systems become less useful for the end-user. As the relationship between intra-speaker's speech variability and speech recognition performance is yet unclear, we began to investigate the nature of this relationship.

Advances in Natural Language Processing
Research in Computing Science 18, 2006, pp. 115-124

A Task Analysis of Nursing Activites Using Spoken Corpora

Hiromi itoh Ozaku[1], Akinori Abe[1], Kaoru Sagara[2], Noriaki Kuwahara[1] and Kiyoshi Kogure[1]

[1] Advanced Telecommunications Research Institute International
{romi,ave,kuwahara,kogure}@atr.jp
[2] Seinan Jo Gakuin University
sagara@seinan-jo.ac.jp

Abstract. This paper illustrates our examination of the "E-nightingale Project," reports the results manually obtained from task analyses of nursing activities using spoken corpora, and describes the possibility of automated task analyses using natural language processing technologies.
Recently, medical malpractice has become a serious social problem and safety measures to prevent and decrease medical malpractice are being taken in some hospitals, for example, by building computerized database in new systems.
The Japanese Nursing Association states in its guidelines that nurses are encouraged to make nursing reports, including medical malpractice reports, and to analyze the cause of accidents, which is helpful for preventing recurrences.
However, it is very difficult for nurses to produce detailed records during their working hours except for malpractice reports. Furthermore, it is hard work for nurses on duty to analyze the recorded data in detail.
As a solution, we have launched the "E-nightingale Project," in which some systems using wearable computers are being developed for the purpose of preventing and reducing medical malpractice.
As part of this project, we built spoken corpora using voice data that were monitored and recorded daily nursing assignments in hospitals with our developed devices. 800 hours of voice data were accumulated, and 70 hours of those were transcribed as the spoken corpora. Then we started analyzing nursing tasks using the spoken corpora, and considered the possibility of automated task analysis using natural language processing technologies.

1 Introduction

Recently, medical malpractice has become a serious social problem [1]. The Japanese Ministry of Health, Labor and Welfare has reported that nursing teams are most frequently involved in medical accidents in hospitals [2]. The Japanese Nursing Association also states in its guidelines that nurses are encouraged to make nursing reports and to analyze the cause of accidents by comparing medical malpractice reports, which is helpful for preventing recurrences. Some nursing

Advances in Natural Language Processing
Research in Computing Science 18, 2006, pp. 125-136

Improved Focused Web Crawling Based on Incremental Q-Learning

Yunming Ye[1], Yan Li[1], Joshua Huang[2] and Xiaofei Xu[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 *IQ-Learning*, a new focused crawling algorithm based on incremental Q-learning. The intuition is to extend previous reinforcement learning based focused crawler with the incremental learning mechanisms so that the system can start with a few initial samples and learns incrementally from the knowledge discovered online. First, a sample detector is used to distill new samples from the crawled Web pages, upon which the page relevance estimator can learn an updated estimation model. Secondly, the updated page relevance information is fed back to the Q value estimator constantly when new pages are crawled so that the Q value estimation model will be improved over time. In this way, the reinforcement learning in focused crawling becomes an incremental process and can be more self-adaptive to tackle the complex Web crawling environments. Comparison experiments have been carried out between the *IQ-Learning* algorithm and other two state-of-the-art focused crawling algorithms. The experimental results show that *IQ-Learning* achieves better performance in most of the target topics.

Keywords: Focused Crawler; Q-Learning; Incremental Learning; Web

1 Introduction

A Web crawler is an information gathering system that traverses the Web by following the hyperlinks from page to page and downloads Web pages that are interested. General Web crawlers visit the Web in an unselective mode. They aim at collecting Web pages as many as possible to build search engines. Different from the general crawler, a focused crawler [1] is an intelligent Web crawler that traverses the Web selectively to download pages in some predefined target topics. Given a search topic and a predefined maximum download number, the goal of a focused crawler is to collect as many relevant pages as possible while retrieving as fewer irrelevant pages as possible in the crawling process.

Focused crawlers are very useful in several Web applications [2], such as collecting Web pages with specific topics for domain-specific search engines, archiving specific page collections for a digital library, and gathering systematic information in some specific topics for market research or survey of literature on the

Advances in Natural Language Processing
Research in Computing Science 18, 2006, pp. 139-150

Automatic Identification of Chinese Stop Words

Feng Zou, Fu Lee Wang, Xiaotie Deng, Song Han

Computer Science Department, City University of Hong Kong
Kowloon Tong, Hong Kong
phenix@cs.cityu.edu.hk, {flwang, csdeng, shan00}@cityu.edu.hk

Abstract. In modern information retrieval systems, effective indexing can be achieved by removal of stop words. Till now many stop word lists have been developed for English language. However, no standard stop word list has been constructed for Chinese language yet. With the fast development of information retrieval in Chinese language, exploring Chinese stop word lists becomes critical. In this paper, to save the time and release the burden of manual stop word selection, we propose an automatic aggregated methodology based on statistical and information models for extraction of the stop word list in Chinese language. The novel algorithm balances various measures and removes the idiosyncrasy of particular statistical measures. Extensive experiments have been conducted on Chinese segmentation for illustration of its effectiveness. Results show that the generated stop word list can improve the accuracy of Chinese segmentation significantly.

1 Introduction

In information retrieval, a document is traditionally indexed by words [10, 11, 17]. Statistical analysis through documents showed that some words have quite low frequency, while some others act just the opposite. For example, words "and", "of", and "the" appear frequently in the documents. The common characteristic of these words is that they carry no significant information to the document. Instead, they are used just because of grammar. We usually refer to this set of words as stop words [10, 11, 21].

Stop words are widely used in many fields. In digital libraries, for instance, elimination of stop words could contribute to reduce the size of the indexing structure considerably and obtain a compression of more than 40% [10]. On the other hand, in information retrieval, removal of stop words could not only help to index effectively, but also help to speed up the calculation and increase the accuracy [20].

Lots of stop word lists have been developed for English language in the past, which are usually based on frequency statistics of a large corpus [21]. The English stop word lists available online [22, 23] are good examples. However, no commonly accepted stop word list has been constructed for Chinese language. Most current researches on Chinese information retrieval make use of manual or simple statistical stop word lists [1, 2, 3], some of which are picked up based on the authors experiences consuming a lot of time. The contents of these stop lists vary a lot from each other. With the fast

Advances in Natural Language Processing
Research in Computing Science 18, 2006, pp. 151-162

Fast Search Algorithm for Sequences of Hierarchically Structured Data

Masaki Murata[1], Masao Utiyama[1], Toshiyuki Kanamaru[1,2] and Hitoshi Isahara[1]

[1] National Institute of Information and Communications Technology,
3-5 Hikaridai, Seika-cho, Soraku-gun, Kyoto, 619-0289, Japan,
{murata,mutiyama,isahara}@nict.go.jp,
WWW home page: http://www.nict.go.jp/jt/a132/members/murata/
[2] Kyoto University,
Yoshida-nihonmatsu-cho, Sakyo-ku, Kyoto, 606-8501, Japan,
kanamaru@hi.h.kyoto-u.ac.jp

Abstract. We developed an algorithm for quickly searching sequences of hierarchically structured data, such as tagged corpora where each word includes information on its part of speech (POS) and minor POS and the word itself. Using our method, we first make a data item where each data item in a lower level is surrounded by two data items in a higher level. We then connect these data items to make a long string and store the string in a database. We use suffix arrays to query the database. Our experiments showed that our method was 194 times faster than a conventional method at fastest and 24 times faster on average. Our method can be used for other kinds of hierarchically structured data, such as Web applications. Methods that can be used on such data are in high demand. For example, our method can be used to retrieve Web text that includes hierarchical information of low, middle, and high semantic levels. If we use our method for such Web text, we can query using the terms, "High semantic level: method", "Word: in", and "Low semantic level: group"; in other words, our retrieval method is more useful and convenient than conventional Web retrieval.

1 Introduction

We developed an algorithm for quickly searching sequences of hierarchically structured data.[3] For example, in the Kyoto Text Corpus [1], each word has information on its part of speech (POS) and minor POS and the word itself. (A minor POS is a more specific POS.) The POS, minor POS, and word can be considered data of the highest layer, the second-highest layer, and the lowest layer. Hierarchically structured data, as explained below, has data from the lowest to the highest layers. The Kyoto Text Corpus has such a structure, so it can be considered to be sequences of hierarchically structured data. The algorithm we propose is for quickly searching sequences of such data. Our algorithm

[3] We obtained a Japanese patent for the algorithm.

© A. Gelbukh (Editor)
Advances in Natural Language Processing
Research in Computing Science 18, 2006, pp. 163-174

Cross-Lingual Question Answering Using Inter Lingual Index Module of EuroWordNet*

Sergio Ferrández and Antonio Ferrández

Natural Language Processing and Information Systems Group
Department of Software and Computing Systems
University of Alicante, Spain
{sferrandez,antonio}@dlsi.ua.es

Abstract. This paper outlines the BRILI cross-lingual English-Spanish-Catalan Question Answering (QA) system. The BRILI is being designed at University of Alicante and will be capable to answer English, Spanish and Catalan questions from English, Spanish and Catalan documents. The starting point is our monolingual Spanish QA system [11] which was presented at the 2005 edition of the Cross-Language Evaluation Forum (CLEF). We describe the extensions to our monolingual QA system that are required, especially the language identification module and the strategy used for the question processing module. The Inter Lingual Index (ILI) Module of EuroWordNet (EWN) is used by the question processing modules. The aim of this is to reduce the negative effect of question translation on the overall accuracy of QA systems.

1 Introduction

The aim of a Question Answering (QA) system is to localize the correct answer to a question in natural language in a non-structured collection of documents, also the situations where the system is not to able to provide an answer should be detected. In the case of a Cross-Lingual QA (CL-QA) system, the question is formulated in a language different from the one of the documents, which increases the difficulty. Nowadays, multilingual QA systems have been recognized as an important issue for the future of Information Retrieval (IR).

In this paper we present BRILI (Spanish acronym for "Question Answering using Inter Lingual Index Module"). It is a CL-QA system for Spanish, English and Catalan. It is designed to localize answers from documents, where both answers and documents are written in the three languages. The system is based on complex pattern matching using NLP tools [1, 4, 7, 12]. Beside, Word Sense Disambiguation (WSD) is applied to improve the system (a new proposal of WSD for nouns based on [2]).

BRILI is fully automatic, including the modules of language identification and question processing. The main goal of this paper is to describe these modules

* This research has been partially funded by the Spanish Government under project CICyT number TIC2003-07158-C04-01 and by the Valencia Government under project number GV04B-268.

Advances in Natural Language Processing
Research in Computing Science 18, 2006, pp. 177-182

Using Terminology and a Concept Hierarchy for Restricted-Domain Question-Answering

Hai Doan-Nguyen and Leila Kosseim

Department of Computer Science and Software Engineering
Concordia University
Montreal, Quebec, H3G 1M8, Canada
haidoan@cse.concordia.ca, kosseim@cse.concordia.ca

Abstract. This paper discusses an important characteristic of restricted-domain question-answering (RDQA): the issue of low precision in finding good answers. We propose methods for improving precision using domain-specific terminology and concept hierarchy to rearrange the candidate list and to better characterize the question-document relevance relationship. Once this relationship has been well established, one can expect to obtain a small set of (almost) all relevant documents for a given question, and use this to guide the information retrieval engine in a two-level search strategy. The methods proposed here have been applied to a real QA system for a large telecommunication company, yielding significant improvements in precision.

1 Introduction

This paper presents our research in the development of a question-answering (QA) system for a restricted domain. The system's goal is to reply to customer's questions on services offered by Bell Canada, a major Canadian telecommunication corporation. Although grounded within a specific context, we try to reveal general problems and develop a general methodology for restricted-domain QA (RDQA).

Although work in RDQA dates back to the early years of Artificial Intelligence, the domain has only recently regained interest in the research community. RDQA performs QA on a specific domain and often uses document collections restricted in subject and volume. Often integrated in real-world applications, RDQA, especially systems working on free text corpora (rather than on structured databases), provides many interesting challenges to natural language engineering. Real questions have no limit on form, style, category, and complexity. In addition, a RDQA system often has to deal with the problem of low precision in finding a correct answer for a given question.

In this paper, we discuss the main characteristics of RDQA, and present a series of methods to improve the precision performance of our system. These methods were not developed specifically for our project at hand but always considered in a general perspective of RDQA.

Advances in Natural Language Processing
Research in Computing Science 18, 2006, pp. 183-194

Efficient Randomized Algorithms for Text Summarization

Ahmed Mohamed and Sanguthevar Rajasekaran

Department of Computer Science & Engineering,
University of Connecticut, Storrs, CT 06268
{amohamed, rajasek}@engr.uconn.edu

Abstract. Text summarization is an important problem since it has numerous applications. This problem has been extensively studied and many approaches have been pro-posed in the literature for its solution. One such interesting approach is that of posing summarization as an optimization problem and using genetic algorithms to solve this optimization problem. In this paper we present elegant randomized algorithms for summarization based on sampling. Our experimental results show that our algorithms yield better accuracy than genetic algorithms while significantly saving on time. We have employed data from Document Understanding Conference 2002 and 2004 (DUC-2002, DUC-2004) in our experiments.

1 Introduction

Document summarization has been the focus of many researchers for the last decade, due to the increase in on-line information and the need to find the most important information in a (set of) document(s). There are different approaches to generate summaries depending on the task the summarization is required for. Summarization approaches usually fall into 3 categories (Mani and Maybury, 1999):

- *Surface-level* approaches tend to represent information in terms of shallow features, which are then selectively combined together to yield a salience function used to extract information;
- *Entity-level* approaches build an internal representation for text, modeling text entities and their relationships. These approaches tend to represent patterns of connectivity in the text (e.g., graph topology to help determine what is salient);
- *Discourse-level* approaches model the global structure of the text, and its relation to communicative goals.

Some approaches mix between two or more of the features of the above mentioned approaches, and the approaches discussed in this paper fall in that category, since they involve both surface and entity levels' features.

© A. Gelbukh (Editor)
Advances in Natural Language Processing
Research in Computing Science 18, 2006, pp. 195-200

Arabic Error Feedback in an Online Arabic Learning System

Khaled F. Shaalan[1,2] and Habib E. Talhami[1,2]

[1] Institute of Informatics
The British University in Dubai
P. O. Box 502216, Dubai,UAE
[2] Honorary Fellow, School of Informatics, University of Edinburgh
{khaled.shaalan,habib.talham}@buid.ac.ae

Abstract. Arabic is a Semitic language that is rich in its morphology and syntax. The very numerous and complex grammar rules of the language could be confusing even for Arabic native speakers. Many Arabic intelligent computer-assisted language-learning (ICALL) systems have neither deep error analysis nor sophisticated error handling. In this paper, we report an attempt at developing an error analyzer and error handler for Arabic as an important part of the Arabic ICALL system. In this system, the learners are encouraged to construct sentences freely in various contexts and are guided to recognize by themselves the errors or inappropriate usage of their language constructs. We used natural language processing (NLP) tools such as a morphological analyzer and a syntax analyzer for error analysis and to give feedback to the learner. Furthermore, we propose a mechanism of correction by the learner, which allows the learner to correct the typed sentence independently. This will result in the learner being able to figure out what the error is. Examples of error analysis and error handling will be given and will illustrate how the system works.

1 Introduction

Computer-assisted language learning (CALL) addresses the use of computers for language teaching and learning. CALL emerged in the early days of computers. Since the early 1960's, CALL systems have been designed and built. The effectiveness of CALL systems has been demonstrated by many researchers [6, 7]. More than a decade ago, Intelligent Computer-Assisted Language Learning (ICALL) started as a separate research field, when artificial intelligence (AI) technologies were mature enough to be included in language learning systems. The beginning of the new research field was characterized by intelligent tutoring systems (ITS), which embedded some NLP features to extend the functionality of traditional language learning systems. The continuous advances in ICALL systems have been documented in several publications [2, 3, 5, 9].

On of the weaknesses of current Arabic ICALL systems is that learners cannot key in Arabic sentences freely. Similarly, the system cannot guide the learner to correct the most likely ill-formed input sentences. The learner just accepts the information,

© *A. Gelbukh (Editor)*
Advances in Natural Language Processing
Research in Computing Science 18, 2006, pp. 203-212

Author Index
Índice de autores

Editorial Board of the Volume

Comité editorial de volumen

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 Reviewers

Árbitros adicionales

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

Impreso en los Talleres Gráficos
de la Dirección de Publicaciones
del Instituto Politécnico Nacional
Tresguerras 27, Centro Histórico, México, D.F.
Febrero de 2006.
Printing 500 / Edición 500 ejemplares.