A Language Independent Approach for Recognizing Textual Entailment

Adrian Iftene and Alexandra Balahur-Dobrescu

"Al.I.Cuza" University of Iasi, Romania
adiftene@info.uaic.ro, alexyys13@yahoo.com

Abstract. Textual Entailment Recognition (RTE) was proposed as a generic task, aimed at building modules capable of capturing the semantic variability of texts and performing natural language inferences. These modules can be then included in any NLP system, improving its performance in fine-grained semantic differentiation. The first part of the article describes our approach aimed at building a generic, language-independent TE system that would eventually be used as a module within a QA system. We evaluated the accuracy of this system by building two instances of it - for English and Romanian and testing them on the data from the RTE3 competition. In the second part we show how we applied the steps described in [1] and adapted this system in order to include it as module in a QA system architecture. Lastly, we show the results obtained, which point out significant growth in precision.

1 Introduction

Recognizing textual entailment RTE[1] [3] is the task of deciding, given two text fragments, whether the meaning of one text is entailed (can be inferred) from the other text. The aim in defining this task was to create an application-independent framework for assessing means of capturing major semantic inferences needed in many NLP applications. Examples of such applications are: Information Retrieval (IR), Question Answering (QA), Information Extraction (IE), and Text Summarization (SUM).

Formally, textual entailment is defined in [1] as a directional relation between two text fragments, termed *T - the entailing text*, and *H - the entailed text*. It is then said that *T* entails *H* if, typically, a human reading *T* would infer that *H* is most likely true. This definition is based on (and assumes) common human understanding of language as well as common Background Knowledge.

TE systems compete each year in the RTE competition, organized by PASCAL[2] (Pattern Analysis, Statistical Modeling and Computational Learning) - the European Commission's IST-funded Network of Excellence for Multimodal Interfaces.

Question Answering (QA) Systems are one of the main research topics in the Natural Language Processing field. These systems not only employ various discourse

[1] http://www.pascal-network.org/Challenges/RTE/
[2] http://www.pascal-network.org/

© A. Gelbukh (Ed.)
Advances in Natural Language Processing and Applications
Research in Computing Science 33, 2008, pp. 3-14

Received 17/10/07
Accepted 07/12/07
Final Version 16/01/08

Summarization by Logic Segmentation and Text Entailment

Doina Tatar, Emma Tamaianu-Morita, Andreea Mihis and Dana Lupsa

University "Babes-Bolyai"
Cluj-Napoca
Romania
{dtatar,mihis,dana}@cs.ubbcluj.ro,etamaian@yahoo.com

Abstract. As the phenomenon of information overload grows day by day, systems that can automatically summarize documents become increasingly studied and used. This paper presents some original methods for text summarization of a single source document by extraction. The methods are based on some of our own text segmentation algorithms. We denote them logical segmentations because for all these methods the score of a sentence is the number of sentences of the text which are entailed by it. The first method of segmentation and summarization is called Logical TextTiling (LTT): for a sequence of sentences, the scores form a structure which indicates how the most important sentences alternate with ones less important and organizes the text according to its logical content. The second method, Pure Entailment uses definition of the relation of entailment between two texts. The third original method applies Dynamic Programming and centering theory to the sentences logically scored as above. The obtained ranked logical segments are used in the summarization. Our methods of segmentation and summarization are applied and evaluated against a manually realized segmentation and summarization of the same text by Donald Richie, "The Koan", [9]. The text is reproduced at [14].

1 Introduction

Systems that can automatically summarize documents become increasingly studied and used. As a summary is a shorter text (usually no longer than a half of the source text) that is produced from one or more original texts keeping the main ideas, the most important task of summarization is to identify the most informative (salient) parts of a text comparatively with the rest. Usually the salient parts are determined on the following assumptions [6]:

- they contain words that are used frequently;
- they contain words that are used in the title and headings;
- they are located at the beginning or end of sections;
- they use key phrases which emphasize the importance in text;
- they are the most highly connected with the other parts of text.

Advances in Natural Language Processing and Applications
Research in Computing Science 33, 2008, pp. 15-26

Received 24/10/07
Accepted 07/12/07
Final Version 12/01/08

TIL as the Logic of Communication in a Multi-Agent System

Marie Duží

VSB-Technical University Ostrava, 17. listopadu 15, 708 33 Ostrava, Czech Republic
marie.duzi@vsb.cz

Abstract. The method of encoding communication of agents in a multi-agent system (MAS) is described. The autonomous agents communicate with each other by exchanging messages formulated in a near-to-natural language. Transparent Intensional Logic (TIL) is an expressive system primarily designed for the logical analysis of natural language; thus we make use of TIL as a tool for encoding the semantic content of messages. The hyper-intensional features of TIL analysis are described, in particular with respect to agents' attitudes and anaphoric references. We demonstrate the power of TIL to determine the antecedent of an anaphoric pronoun. By an example of a simple dialogue we illustrate the way TIL can function as a dynamic logic of discourse where anaphoric pronouns refer to entities of any type, even constructions, i.e. the structured meanings of other expressions.

1 Introduction

Multi-agent system (MAS) is a system composed of autonomous, intelligent but resource-bounded agents. The agents are active in their perceiving environment and acting in order to achieve their individual as well as collective goals. They communicate with each other by exchanging messages formulated in a standardised natural language. According to the FIPA standards[1], *message* is the basic unit of communication. It can be of an arbitrary form but it is supposed to have a structure containing several attributes. Message semantic *content* is one of these attributes, the other being for instance 'Performatives', like 'Query', 'Inform', 'Request' or 'Reply'. The content can be encoded in any suitable language. The standards like FIPA SL and KIF are mostly based on the First-Order Logic (FOL) paradigm, enriched with higher-order constructs wherever needed.[2] The enrichments extending FOL are well defined syntactically, while their semantics is often rather sketchy, which may lead to communication inconsistencies. Moreover, the bottom-up development from FOL to more complicated cases yields the versions that do not fully meet the needs of the MAS communication. In particular, agents' attitudes and anaphora processing create a problem. In the paper we are going to demonstrate the need for an expressive logical tool of Transparent Intensional Logic (TIL) in order to encode the semantic content of

[1] The Foundation for Intelligent Physical Agents, http://www.fipa.org

[2] For details on FIPA SL, see http://www.fipa.org/specs/fipa00008/; for KIF, Knowledge Interchange Format, see http://www-ksl.stanford.edu/knowledge-sharing/kif/

Advances in Natural Language Processing and Applications
Research in Computing Science 33, 2008, pp. 27-20

Received 18/10/07
Accepted 07/12/07
Final Version 12/01/08

Collins-LA: Collins' Head-Driven Model with Latent Annotation

Seung-Hoon Na[1] , Meixun Jin[1], In-Su Kang[2] and Jong-Hyeok Lee[1]

[1] Pohang University of Science and Technology (POSTECH), AITrc, Republic of Korea
{nsh1979,meixunj,jhlee}@postech.ac.kr,
[2] Korea Institute of Science and Technology Information (KISTI), Republic of Korea
dbaisk@kisti.re.kr

Abstract. Recent works on parsing have reported that the lexicalization does not have a serious role for parsing accuracy. Latent-annotation methods such as PCFG-LA are one of the most promising un-lexicalized approaches, and reached the-state-of-art performance. However, most works on latent annotation have investigated only PCFG formalism, without considering the Collins' popular head-driven model, though it is a significantly important and interesting issue. To this end, this paper develops Collins-LA, the extension of the Collins' head-driven model to support the latent annotation. We report its basic accuracy, comparing with PCFG-LA. The experimental results show that Collins-LA has potential to improve basic parsing accuracy, resulting in comparable performance with PCFG-LA even in the naive setting.

1 Introduction

Recent works for parsing have consistently shown that the lexicalization does not have serious effects on parsing accuracy. Gildea mentioned that the high performance of a Collins' model is obtained not from the bi-lexical dependency, showing that parsing accuracy is not decreased even when the bi-lexical dependency is not incorporated to the model [1]. Gildea's result has been re-confirmed by Bikel, during his investigation through the re-implementation of the Collins' parsing model [2].

Another direction is opened from a Klein's work, where fully un-lexicalized parsing models are extensively evaluated through an accurate design of tree-bank annotation [3]. Klein's work includes Johnson's parent annotation [4], and external-internal annotation, tag-splitting, and head annotation, etc, resulting in the parsing accuracy of about 86%, which corresponds to the initial performance of the lexicalized parsing accuracy. Motivated from this, Matsuzaki proposed a new generative model PCFG-LA, an extension of PCFG models in which non-terminal symbols are annotated with latent variables [5]. PCFG-LA successfully replaces the manual feature selection used in previous research such as Johnson's work or Klein's work, by showing that PCFG-LA reaches Klein's result. Based on this, Petrov et al. further explored PCFG-LA where hierarchical splitting and merging of latent non-terminals are utilized, thus differentiating the number of latent variables of each non-terminal symbol [6]. Their result is remarkable, showing about 90% accuracy, which is almost comparable to the-state-of-art of Charniak's parser [7]. All these previous results consistently show

Advances in Natural Language Processing and Applications
Research in Computing Science 33, 2008, pp. 43-54

Received 15/10/07
Accepted 07/12/07
Final Version 19/01/08

Study on Architectures for Chinese POS Tagging and Parsing

Hailong Cao, Yujie Zhang and Hitoshi Isahara

3-5 Hikari-dai, Seika-cho, Soraku-gun, Kyoto, 619-0289, Japan
{hlcao, yujie, isahara}@nict.go.jp

Abstract. How to deal with part of speech (POS) tagging is a very important problem when we build a syntactic parsing system. We could preprocess the text with a POS tagger before perform parsing in a pipelined approach. Alternatively, we could perform POS tagging and parsing simultaneously in an integrated approach. Few, if any, comparisons have been made on such architecture issues for Chinese parsing. This paper presents an in-depth study on this problem. According to comparison experiments, we find that integrated approach can make significantly better performance both on Chinese parsing and unknown words POS tagging than the pipelined approach. As for known words POS tagging, we find that the two approaches get similar tagging accuracy, but the tagging results of integrated approach do lead to much better parsing performance. We also analyze the reasons account for the performance difference.

1 Introduction

POS tag is an important feature in most of the parsing models as having a word's POS tag can help us determine what kind of syntactic constituent the word can compose. So usually it is necessary to assign a proper POS tag to each word in a sentence which is to be parsed. We could adopt the pipelined approach which performs parsing strictly after POS tagging, or performs POS tagging and parsing simultaneously in an integrated approach. The pipelined approach is simple and fast but is subject to error propagation. Though integrated approach can make decision from global view in theory, whether it can get better accuracy in practice is still an open question since little detailed comparison has been made between pipelined and integrated approaches for Chinese parsing.

This paper presents an in-depth study on such issues for Chinese parsing. We compare the performances of the pipelined approach, the integrated approach and two kinds of compromise strategies. There are three findings in our experiments. First, integrated approach can improve parsing performance by considering POS tag of known word globally though it can not enhance the known words tagging accuracy. Second, integrated approach can get better tagging accuracy on unknown words and therefore get better parsing result. Third, better tagging results do not always lead to better parsing results. Our comparison experiments suggest that fully integrated approach is the best strategy for Chinese parsing if complexity is not a major concern. We also analyze the reasons that account for the performance difference.

Advances in Natural Language Processing and Applications
Research in Computing Science 33, 2008, pp. 55-66

Received 11/10/07
Accepted 07/12/07
Final Version 17/01/08

Maximum Entropy Based Bengali Part of Speech Tagging

Asif Ekbal[1], Rejwanul Haque[2], and Sivaji Bandyopadhyay[3]

Computer Science and Engineering Department, Jadavpur University, Kolkata, India
asif.ekbal@gmail.com[1], rejwanul@gmail.com[2], sivaji_cse_ju@yahoo.com[3]

Abstract. Part of Speech (POS) tagging can be described as a task of doing automatic annotation of syntactic categories for each word in a text document. This paper presents a POS tagger for Bengali using the statistical Maximum Entropy (ME) model. The system makes use of the different contextual information of the words along with the variety of features that are helpful in predicting the various POS classes. The POS tagger has been trained with a training corpus of 72, 341 word forms and it uses a tagset [1] of 26 different POS tags, defined for the Indian languages. A part of this corpus has been selected as the development set in order to find out the best set of features for POS tagging in Bengali. The POS tagger has demonstrated an accuracy of 88.2% for a test set of 20K word forms. It has been experimentally verified that the lexicon, named entity recognizer and different word suffixes are effective in handling the unknown word problems and improve the accuracy of the POS tagger significantly. Performance of this system has been compared with a Hidden Markov Model (HMM) based POS tagger and it has been shown that the proposed ME based POS tagger outperforms the HMM based tagger.

Keywords: Part of Speech Tagging, Maximum Entropy Model, Bengali.

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. Part of speech tagging is a very important preprocessing task for language processing activities. This helps in doing deep parsing of text and in developing Information extraction systems, semantic processing etc. Part of speech tagging for natural language texts are developed using linguistic rules, stochastic models and a combination of both. Stochastic models [1] [2] [3] have been widely used in POS tagging task for simplicity and language independence of the models. Among stochastic models, Hidden Markov Models (HMMs) are quite popular. Development of a stochastic tagger requires large amount of annotated corpus. Stochastic taggers with more than 95% word-level accuracy have been developed for English,

[1] http://shiva.iiit.ac.in/SPSAL2007/iiit_tagset_guidelines.pdf

Advances in Natural Language Processing and Applications
Research in Computing Science 33, 2008, pp. 67-78

Received 28/10/07
Accepted 07/12/07
Final Version 18/01/08

A Two-Pass Search Algorithm for Thai Morphological Analysis

Canasai Kruengkrai and Hitoshi Isahara

Graduate School of Engineering, Kobe University
1-1 Rokkodai-cho, Nada-ku, Kobe 657-8501 Japan
National Institute of Information and Communications Technology
3-5 Hikaridai, Seika-cho, Soraku-gun, Kyoto 619-0289 Japan
{canasai,isahara}@nict.go.jp

Abstract. Considering Thai morphological analysis as a search problem, the approach is to search the most likely path out of all candidate paths in the word lattice. However, the search space may not contain all possible word hypotheses due to the unknown word problem. This paper describes an efficient algorithm called the *two-pass search algorithm* that first recovers missing word hypotheses, and then searches the most likely path in the expanded search space. Experimental results show that the two-pass search algorithm improves the performance of the standard search by 3.23 F_1 in word segmentation and 2.92 F_1 in the combination of word segmentation and POS tagging.

1 Introduction

Morphological analysis has been recognized as a fundamental process in Thai text analysis. In Thai, words are written continuously without word boundaries like other non-segmenting languages (e.g., Chinese and Japanese). However, the Thai writing system has certain unique characteristics. For example, in order to form a smallest linguistic unit, a character cluster, including a consonant, a vowel, and/or a tonal mark, must be formed.

Thai morphological analysis generally involves two tasks: segmenting a character string into meaningful words, and assigning words with the most likely part-of-speech (POS) tags. Considering Thai morphological analysis as a search problem, the approach is to search the most likely path out of all candidate paths in a word lattice. Figure 1 illustrates the word lattice of a string analysis, consisting of possible word hypotheses and their connections. The path indicated by the bold line is the correct segmentation.

Discriminating such path from other candidate paths is a difficult problem itself, and requires a dynamic programming search (e.g., the Viterbi algorithm). The problem becomes more difficult, since the search space is often imperfect. Some word hypotheses are missing due to the unknown word problem. Unknown words are words that do not occur in the system dictionary or the training corpus. The system has no knowledge to use in generating hypotheses for unknown words.

Advances in Natural Language Processing and Applications
Research in Computing Science 33, 2008, pp. 81-92

Received 15/10/07
Accepted 07/12/07
Final Version 21/01/08

Exploiting Unlabeled Text with Different Unsupervised Segmentation Criteria for Chinese Word Segmentation

Hai Zhao and Chunyu Kit

Department of Chinese, Translation and Linguistics,
City University of Hong Kong,
83 Tat Chee Avenue, Kowloon, Hong Kong, China
{haizhao,ctckit}@cityu.edu.hk

Abstract. This paper presents a novel approach to improve Chinese word segmentation (CWS) that attempts to utilize unlabeled data such as training and test data without annotation for further enhancement of the state-of-the-art performance of supervised learning. The lexical information plays the role of information transformation from unlabeled text to supervised learning model. Four types of unsupervised segmentation criteria are used for word candidate extraction and the corresponding word likelihood computation. The information output by unsupervised segmentation criteria as features therefore is integrated into supervised learning model to strengthen the learning for the matching subsequence. The effectiveness of the proposed method is verified in data sets from the latest international CWS evaluation. Our experimental results show that character-based conditional random fields framework can effectively make use of such information from unlabeled data for performance enhancement on top of the best existing results.

1 Introduction

The task of Chinese word segmentation (CWS) is to segment an input sequence of characters into a sequence of words. It is also a preprocessing task shared by many Asian languages without overt word delimiters. CWS was first formulated as a character tagging problem in [1], via labeling each character's position in a word. For example, the segmentation for following sentences,

他 / 来自 / 墨西哥 。
(he / comes from / Mexico.),

receives the tag (label) sequence $SBEBME$ as segmentation result, where the four tags B, M and E stand for the beginning, middle and ending positions in a word, and S for a single character as a word. A Maximum Entropy (MaxEnt) model was trained for such a tagging task in [1]. Many supervised learning methods have been successfully applied to CWS since the First International Chinese Word Segmentation Bakeoff in 2003 [2]. Among them, the character tagging is a particularly simple but effective formulation of the problem suitable for various competitive supervised machine learning models such as MaxEnt, conditional random fields (CRFs), and support vector machines. [1, 3–8].

Advances in Natural Language Processing and Applications
Research in Computing Science 33, 2008, pp. 93-104

Received 07/10/07
Accepted 07/12/07
Final Version 17/01/08

Defining Relative Strengths of Classifiers at Randomly Generated Datasets

Harri M.T. Saarikoski

Department of Translation Studies
Helsinki University, Finland
Harri.Saarikoski@helsinki.fi

Abstract. Classification is notoriously computing-intensive, especially with large datasets. The common (mis)understanding is that cross-validation (CV) tests on subsets of the training data are the only recourse for the selection of best classifier for given dataset. We have shown in [9] that best classifier can be selected on basis of a few prediction factors (related to training volume, number of features and feature value distribution) calculated from the dataset. In this paper, we report promising results of a series of runs with a number of strong classifiers (Support Vector Machine kernels, Decision Tree variants, Naive Bayes and also K Nearest Neighbor Classifiers and Neural Networks) with randomly generated datasets (gensets). We outline the relative strengths of these classifiers in terms of four prediction factors and generalize the findings to different types of datasets (word sense disambiguation, gensets and a protein recognition dataset).

1 Introduction

Classification is the process of automatically resolving the conflicts that occur when two or more different things have the same name and are thereby ambiguous to machines (e.g. the English noun bank can refer among other things to a financial institution or a river embankment). It is a required procedure for machine-learning application fields as varied as data mining (e.g. recognition of the protein type), text mining and natural language processing (NLP). For instance, word sense disambiguation (WSD) is defined as the resolution of ambiguous words into meanings or senses). To resolve these 'classification tasks' automatically, training instances containing independent observations (instances) of ambiguous classes are gathered (or extracted from text in case of WSD). Then from this training data, a training model is learned by classifiers, and the model is tested or evaluated against unseen (i.e. test) instances. Classification accuracy (number of correctly classified test instances / total number of test instances) is the qualitative measure used in this paper for assessing the performance of classifiers.

In the next sections, we describe the investigated datasets, prediction factors and classifier differences. Then we present the summary of results from those classifiers at those datasets and explain the systematic findings in terms of those prediction factors and finally conclude on the importance of the findings.

Advances in Natural Language Processing and Applications
Research in Computing Science 33, 2008, pp. 107-120

Received 28/10/07
Accepted 07/12/07
Final Version 18/01/08

Features and Categories Design for the English-Russian Transfer Model

Elena Kozerenko
Institute for Informatics Problems of the Russian Academy of Sciences,
44 Corpus 2, Vavilova Str., 119333, Moscow, Russia
elenakozerenko@yahoo.com

Abstract. The paper focuses on the role of features for the implementation of the transfer-based machine translation systems. The semantic content of syntactic structures is established via the contrastive study of the English and Russian language systems and parallel texts analysis. The notion of cognitive transfer is employed which means that a language unit or structure can be singled out for transfer when there exists at least one language unit or structure with a similar meaning in the target language The approach taken is aimed at providing computational tractability and portability of linguistic presentation solutions for various language engineering purposes.

Keywords: Machine translation, syntactic structures, features, categories, cognitive transfer.

1 Introduction

The rapid development of language processing systems within the statistical frameworks has revealed the limitations of purely statistical methods in machine translation projects, and at the same time stimulated the new approaches to linguistic rule systems design making them adjusted to be used together with the modern statistical methods. The rule system described in this paper builds on the awareness of the fact that the meaning of a structure in a source language may shift to another category in the language of translation. This awareness is very important for obtaining reliable statistical information from parallel texts corpora to be further used in statistical machine translation algorithms. Otherwise the existing stochastic methods for language processing bring a lot of excessive inconsistent rules which still require filtering and hand editing.

Generally, major efforts connected with natural language modeling lay emphasis at lexical semantics presentations and less attention is paid to the semantics of structures and establishment of functional similarity of language patterns as a core problem in multilingual systems design. The studies presented in this paper focus on the semantics of language structures, namely, the interaction of categorial and functional meanings for subsequent language engineering design of feature-value structures. The proposed methods of dealing with syntactic synonymy of structures (isofunctionality) and structural (syntactic) polysemy provide an essential linguistic foundation for learning mechanisms.

Advances in Natural Language Processing and Applications
Research in Computing Science 33, 2008, pp. 123-138
Received 14/10/07
Accepted 07/12/07
Final Version 15/01/08

Translation of the Light Verb Constructions in Japanese-Chinese Machine Translation

Yiou Wang[1] and Takashi Ikeda[2]

[1]Graduate school of Engineering, Gifu University
y_wang@ikd.info.gifu-u.ac.jp
[2]Yanagido 1-1, Gifu, Gifu 501-1193, Japan
ikeda@info.gifu-u.ac.jp

Abstract. We study the light verb constructions in Japanese, the constructions of expressions with light verb "suru". Such factors as numerous semantic variants and the syntactic complexity of corresponding Chinese expressions of LVCs put great difficulties in the Japanese-Chinese translation, so current commercially available Japanese-Chinese translation softwares give bad results for LVCs translation. We systematically analyze and describe the phenomena and propose translation rules for the conversion of Japanese LVCs into Chinese. we conducted manual experiments using 200 sentences to evaluate these rules and achieved a translation accuracy of over 80%. We also introduce jaw, a pattern-based translation engine, which can be applied to any target language. We implemented our translation rules for LVCs in the Chinese version, jaw/Chinese and verified their applicability in real MT system by experiments using 100 sentences. Evaluations both by hand and by the machine indicate that it provides high performance and utility.

1 Introduction

In order to examine light verb constructions we must first consider the nature of light verbs. The term "light verb" first occurs in Jespersen [1] and Cattell [2].A light verb is a verb (usually a frequent verb with a very general meaning) which, when combined with certain complements, loses some of its normal semantics: such usages are said to be *semantically bleached*[3].Light verb constructions (hereafter LVCs) is a multiword expression that combines a light verb with a complement of noun, adjective, preposition or verb etc. [4].Light verbs are a cross-linguistic phenomenon, and are found in languages such as Persian, Japanese, and Chinese [5][6].In English the most usual light verbs are "make" "do" "take" "have" and "give" as in such LVCs as "have a look" "take a rest" "do a play" "give a hug" and "make a telephone call". In Grimshaw and Mester[7], the Japanese verb *suru* 'do' is analyzed as a light verb. Like the English light verbs, *suru* is assumed to be thematically "light" or incomplete for the contribution to the meaning of verb phrases. We found many studies about the Japanese LVCs (e.g., [8][9][10][11]). These studies discuss Japanese LVCs from various viewpoints and also define different scopes for Japanese LVCs. In this paper we use the terminology "LVC" in the broadest sense: A construction is a LVC if the light verb *suru* is involved in. It maybe questionable whether some sort of constructions we consider here can still be called light verb constructions, but we also investigate such sort of constructions for the sake of completeness.

Advances in Natural Language Processing and Applications
Research in Computing Science 33, 2008, pp. 139-150

Received 12/10/07
Accepted 07/12/07
Final Version 16/01/08

Shahmukhi to Gurmukhi Transliteration System: A Corpus based Approach

Tejinder Singh Saini[1] and Gurpreet Singh Lehal[2]

[1] Advanced Centre for Technical Development of Punjabi Language, Literature & Culture, Punjabi University, Patiala 147 002, Punjab, India
tej@pbi.ac.in
http://www.advancedcentrepunjabi.org
[2] Department of Computer Science, Punjabi University, Patiala 147 002, Punjab, India
gslehal@yahoo.com

Abstract. This research paper describes a corpus based transliteration system for Punjabi language. The existence of two scripts for Punjabi language has created a script barrier between the Punjabi literature written in India and in Pakistan. This research project has developed a new system for the first time of its kind for Shahmukhi script of Punjabi language. The proposed system for Shahmukhi to Gurmukhi transliteration has been implemented with various research techniques based on language corpus. The corpus analysis program has been run on both Shahmukhi and Gurmukhi corpora for generating statistical data for different types like character, word and n-gram frequencies. This statistical analysis is used in different phases of transliteration. Potentially, all members of the substantial Punjabi community will benefit vastly from this transliteration system.

1 Introduction

One of the great challenges before Information Technology is to overcome language barriers dividing the mankind so that everyone can communicate with everyone else on the planet in real time. South Asia is one of those unique parts of the world where a single language is written in different scripts. This is the case, for example, with Punjabi language spoken by tens of millions of people but written in Indian East Punjab (20 million) in Gurmukhi script (*a left to right script based on Devanagari*) and in Pakistani West Punjab (80 million), written in Shahmukhi script (*a right to left script based on Arabic*), and by a growing number of Punjabis (2 million) in the EU and the US in the Roman script. While in speech Punjabi spoken in the Eastern and the Western parts is mutually comprehensible, in the written form it is not. The existence of two scripts for Punjabi has created a script barrier between the Punjabi literature written in India and that in Pakistan. More than 60 per cent of Punjabi literature of the medieval period (500-1450 AD) is available in Shahmukhi script only, while most of the modern Punjabi writings are in Gurmukhi. Potentially, all members of the substantial Punjabi community will benefit vastly from the transliteration system.

Advances in Natural Language Processing and Applications
Research in Computing Science 33, 2008, pp. 151-162

Received 25/10/07
Accepted 07/12/07
Final Version 22/01/08

Vector based Approaches to Semantic Similarity Measures

Juan M. Huerta

IBM T. J. Watson Research Center
1101 Kitchawan Road, Yorktown Heights, NY, 10598
huerta@us.ibm.com

Abstract. This paper describes our approach to developing novel vector based measures of semantic similarity between a pair of sentences or utterances. Measures of this nature are useful not only in evaluating machine translation output, but also in other language understanding and information retrieval applications. We first describe the general family of existing vector based approaches to evaluating semantic similarity and their general properties. We illustrate how this family can be extended by means of discriminatively trained semantic feature weights. Finally, we explore the problem of rephrasing (i.e., addressing the question *is sentence X the rephrase of sentence Y?*) and present a new measure of the semantic linear equivalence between two sentences by means of a modified LSI approach based on the Generalized Singular Value Decomposition.

1 Introduction

Measurements of semantic similarity between a pair of sentences[1] provide a fundamental function in NLU, machine translation, information retrieval and voice based automation tasks, among many other applications. In machine translation, for example, one would like to quantitatively measure the quality of the translation output by measuring the effect that translation had in the conveyed message. In voice based automation tasks, for example in natural language call routing applications, one approach one could take is to compare the uttered input against a collection of canonical or template commands deeming the closest category as the intended target.

Current approaches to semantic similarity measurement include techniques that are specific or custom to the task at hand. For example, in machine translation, the BLEU metric [1] is used in measuring similarity of the MT output. In call routing, vector based methods (e.g., [2, 3]) are used to compare the input utterance against a set of template categories. In information retrieval some approaches use the cosine distance between a query and a document-vector mapped into a lower dimension LSI concept

[1] In this paper, for the sake of conciseness, we use the terms *document*, *utterance*, and *sentence* interchangeably. Typically the nature of the task define the specific type (for example, voice automation systems use *utterances* and so on).

Advances in Natural Language Processing and Applications
Research in Computing Science 33, 2008, pp. 163-174

Received 15/10/07
Accepted 07/12/07
Final Version 29/01/08

Comparing and Combining Methods for Automatic Query Expansion *

José R. Pérez-Agüera[1] and Lourdes Araujo[2]
jose.aguera@fdi.ucm.es, lurdes@lsi.uned.es

[1]Dpto. de Ingeniería del Software e Inteligencia Artificial, UCM, Madrid 28040, Spain,
[2]Dpto. Lenguajes y Sistemas Informáticos. UNED, Madrid 28040, Spain,

Abstract. Query expansion is a well known method to improve the performance of information retrieval systems. In this work we have tested different approaches to extract the candidate query terms from the top ranked documents returned by the first-pass retrieval. One of them is the cooccurrence approach, based on measures of cooccurrence of the candidate and the query terms in the retrieved documents. The other one, the probabilistic approach, is based on the probability distribution of terms in the collection and in the top ranked set. We compare the retrieval improvement achieved by expanding the query with terms obtained with different methods belonging to both approaches. Besides, we have developed a naïve combination of both kinds of method, with which we have obtained results that improve those obtained with any of them separately. This result confirms that the information provided by each approach is of a different nature and, therefore, can be used in a combined manner.

1 Introduction

Reformulation of the user queries is a common technique in information retrieval to cover the gap between the original user query and his need of information. The most used technique for query reformulation is query expansion, where the original user query is expanded with new terms extracted from different sources. Queries submitted by users are usually very short and query expansion can complete the information need of the users.

A very complete review on the classical techniques of query expansion was done by Efthimiadis [5]. The main problem of query expansion is that in some cases the expansion process worsens the query performance. Improving the robustness of query expansion has been the goal of many researchers in the last years, and most proposed approaches use external collections [17, 16, 15], such as the Web documents, to extract candidate terms for the expansion. There are other methods to extract the candidate terms from the same collection that

* Supported by project TIC2007-67581-C02-01/ and Dirección General de Universidades e Investigación de la Consejería de Educación de la Comunidad de Madrid and Universidad Complutense de Madrid (Grupo de investigación consolidado 910494)

Advances in Natural Language Processing and Applications
Research in Computing Science 33, 2008, pp. 177-188

Received 15/10/07
Accepted 07/12/07
Final Version 17/01/08

Meta-Search Utilizing Evolutionary Recommendation: A Web Search Architecture Proposal

Dušan Húsek[1], Keyhanipour[2], Pavel Krömer[3], Behzad Moshiri[2], Suhail Owais[4], Václav Snášel[3]

[1] Institute of Computer Science of Academy of Sciences of the Czech Republic, 18207 Prague, Czech Republic
dusan@cs.cas.cz
[2] Control and Intelligent Processing Center of Excellence, School of Electrical and Computer Engineering, University of Tehran, Iran
a.keyhanipour@ieee.org, moshiri@ut.ac.ir
[3] Department of Computer Science, Faculty of Electrical Engineering and Computer Science, VŠB Technical University of Ostrava, 17. listopadu 15, 708 33 Ostrava Poruba, Czech Republic
{pavel.kromer.fei, vaclav.snasel}@vsb.cz
[4] Information Technology, Al-Balqa Applied University - Ajloun University College, P.O. Box 6, JO 26810 Ajloun, Jordan
suhailowais@yahoo.com

Abstract. An innovative meta-search engine named WebFusion has been presented. The search system learns the expertness of every particular underlying search engine in a certain category based on the users preferences according to an analysis of click-through behavior. In addition, an intelligent re-ranking method based on ordered weighted averaging (OWA) was introduced. The re-ranking method was used to fuse the results scores of the underlying search engines. Independently, a progressive application of evolutionary computing to optimize Boolean search queries in crisp and fuzzy information retrieval systems was investigated, evaluated in laboratory environment and presented. In this paper we propose an incorporation of these two innovative recent methods founding an advanced Internet search application.

1 Motivation

WWW consists of more than ten billion publicly visible web documents [1] distributed on millions of servers world-wide. It is a fast growing and continuously changing dynamic environment. Individual general-purpose search engines providing consensual search services have been unable to keep up with this growth. The coverage of the Web by each of the major search engines has been steadily decreasing despite their effort to comprehend larger porting of web space. Several investigations show that no single standalone search engine has complete coverage and it is unlikely any single web search engine ever will [2]. Rather

Advances in Natural Language Processing and Applications
Research in Computing Science 33, 2008, pp. 189-200

Received 14/10/07
Accepted 07/12/07
Final Version 17/01/08

Structuring Job Search via Local Grammars

Sandra Bsiri[1], Michaela Geierhos[1], and Christoph Ringlstetter[2]

[1] CIS, University of Munich
[2] AICML, University of Alberta

Abstract. The standard approach of job search engines disregards the structural aspect of job announcements in the Web. Bag-of-words indexing leads to a high amount of noise. In this paper we describe a method that uses local grammars to transform unstructured Web pages into structured forms. Evaluation experiments show high efficiency of information access to the generated documents.

1 Introduction

After years of steady growth, the main source of information about job announcements is the Internet [1]. Though, there is some bastion left for high profile openings and local jobs, the traditional newspaper advertisement is of declining importance. Big organizations such as corporations or universities provide an obligatory *career link* on their home pages that leads to their job openings. According to a recent study [2], for example, 70% of the workforce in France searches for jobs on the Internet.[3] The interface arranging access to the information on job opportunities is provided by specialized job search engines. Due to the sheer amount of data, a sophisticated technology which guarantees relevancy would be required. In reality, though, search results are rife with noise.

As compared to a standard search engine, job engines are specialized in that they only index a certain part of the document space: pages that contain job announcements. Unfortunately, in most cases, at this point the specialized treatment of the data has its end. The majority of engines uses variants of the standard vector space model [3] to built up an index that later on is approached by similarity based query processing. This blind statistical model leads often to poor results caused, for example, by homography relations between job descriptors and proper names: in a German language environment a search for a position as a *restaurant chef* (German: *Koch*) easily leads to a contamination with documents that mention a "Mr. Koch" from human resources, with "Koch" being a popular German name. Problems like these arise because the used bag-of-words model treats all terms equally without being aware of their context.

The starting point of a solution is the structured nature of the data spanning the search space and the queries accessing it. Unlike a general search scenario, job search can be seen as a slot-filling process. The indexing task is then to detect concept-value pairs in the HTML-documents and make them accessible. Other

[3] For certain groups such as IT-professionals this value probably comes close to 100%.

Advances in Natural Language Processing and Applications
Research in Computing Science 33, 2008, pp. 201-212

Received 25/10/07
Accepted 07/12/07
Final Version 16/01/08

ExpLSA: An Approach Based on Syntactic Knowledge in Order to Improve LSA for a Conceptual Classification Task

Nicolas Béchet and Mathieu Roche and Jacques Chauché

LIRMM - UMR 5506 - CNRS, Univ. Montpellier 2,
34392 Montpellier Cedex 5 - France

Abstract. Latent Semantic Analysis (LSA) is nowadays used in various thematic like cognitive models, educational applications but also in classification. We propose in this paper to study different methods of proximity of terms based on LSA. We improve this semantic analysis with additional semantic information using Tree-tagger or a syntactic analysis to expand the studied corpus. We finally apply LSA on the new expanded corpus.

1 Introduction

Classification's domain has many research fields like conceptual classification. This one consists in gathering terms in concepts defined by an expert. For example, *exhaust pipe, windshield wiper*, and *rearview mirror* terms can be associated to the *automobile* concept. Then, these terms are classified by semantic proximity with different algorithms like k nearest neighbor (KNN) or k means. The corpora have different types as the language, the syntax, the domain (biology, medicine, etc) using a specialized semantic, etc. Then these complex textual data require a specific process.

In this paper, we describe the first step of a conceptual classification, the study of proximity of the terms. First, we use the Latent Semantic Analysis (LSA) method evolved by [1][1]. LSA is a statistic method applied to high dimension corpora to gather terms (conceptual classification) or contexts (textual classification). After the latent semantic analysis application on the corpus, a semantic space associating each word to a vector is returned. Then, the proximity of the two words can be obtained by measuring the cosine between two vectors. Our aim is to improve the performance of the LSA method by an approach named *ExpLSA*.

The *ExpLSA* approach (context **Exp**ansion with **LSA**) consists in expanding the corpus before the application of a "traditional" latent semantic analysis. This context expansion uses semantic knowledge obtained by syntax, what allows to use *ExpLSA* as well specialized corpus as not. Actually, it is not necessary to use training corpus, so it is not necessary to know the general domain of the corpus.

[1] http://www.msci.memphis.edu/∼wiemerhp/trg/lsa-followup.html

Advances in Natural Language Processing and Applications
Research in Computing Science 33, 2008, pp. 213-224

Received 21/10/07
Accepted 07/12/07
Final Version 11/01/08

Gathering Definition Answers by Information Gain

Carmen Martínez-Gil[1] and A. López-López[1]

[1] Instituto Nacional de Astrofísica, Óptica y Electrónica,
Luis Enrique Erro #1, Santa María Tonanzintla, Puebla, 72840, México
{Carmen, allopez}@ccc.inaoep.mx

Abstract. A definition question is a kind of question whose answer is a complementary set of sentence fragments called nuggets, which define the target term. Since developing general and flexible patterns with a wide coverage to answer definition questions is not feasible, we propose a method using information gain to retrieve the most relevant information. To obtain the relevant sentences, we compared the output of two retrieval systems: JIRS and Lucene. One important feature that impacts on the performance of definition question answering systems is the length of the sentence fragments, so we applied a parser to analyze the relevant sentences in order to get clauses. Finally, we observed that, in most of the clauses, only one part before and after the target term contains information that defines the term, so we analyzed separately the sentence fragments before (*left*) and after (*right*) the target term. We performed different experiments with the collections of questions from the *pilot* evaluation of definition questions 2002, *definition* questions from TREC 2003 and *other* questions from TREC 2004. F-measures obtained are competitive when compared against the participating systems in their respective conferences. Also the best results are obtained with the general purpose system (Lucene) instead of JIRS, which is intended to retrieve passages for factoid questions.

1 Introduction

Question Answering (QA) is a computer-based task that tries to improve the output generated by Information Retrieval (IR) systems. A definition question [9] is a kind of question whose answer is a complementary set of sentence fragments called nuggets.

After identifying the correct target term (the term to define) and context terms, we need to obtain useful and non redundant definition nuggets. Nowadays, patterns are obtained manually as surface patterns [5, 6, 12]. These patterns can be very rigid, leading to the alternative soft patterns [2], which are even extracted in an automatic way [5]. Then, once we have the patterns, we apply a matching process to extract the nuggets. Finally, we need to perform a process to determine if these nuggets are part of the definition; where a common criterion employed is the frequency of appearance of the nugget.

According to the state of the art, the highest F-measure in a pilot evaluation [9] for definition questions in 2002 is 0.688 using the nugget set supplied by author, taking

Advances in Natural Language Processing and Applications
Research in Computing Science 33, 2008, pp. 225-234

Received 15/10/07
Accepted 07/12/07
Final Version 11/01/08

Improving Word-Based Predictive Text Entry with Transformation-Based Learning

David J. Brooks and Mark G. Lee

School of Computer Science
University of Birmingham
Birmingham, B15 2TT, UK
d.j.brooks, m.g.lee@cs.bham.ac.uk

Abstract. Predictive text interfaces allow for text entry on mobile phones, using only a 12-key numeric keypad. While current predictive text systems require only around 1 keystroke per character entered, there is ambiguity in the proposed words that must be resolved by the user. In word-based approaches to predictive text, a user enters a sequence of keystrokes and is presented with an ordered list of word proposals. The aim is to minimise the number of times a user has to cycle through incorrect proposals to reach their intended word.
This paper considers how contextual information can be incorporated into this prediction process, while remaining viable for current mobile phone technology. Our hypothesis is that Transformation-Based Learning is a natural choice for inducing predictive text systems. We show that such a system can: a) outperform current baselines; and b) correct common prediction errors such as "of" vs. "me" and "go" vs. "in".

1 Introduction

Since the widespread adoption of the Short Message Service (SMS) protocol, "text-messaging" has become a primary communication medium, and the need for efficient text-entry interfaces on mobile phone keypads has become paramount. The T9 – or "text on 9 keys" – interface allows users to enter alphabetic characters using only 9 keys. Each of these 9 keys is mapped to a set of alphabetic characters according to international standard ITU-T Recommendation E.161 [1, p.1], shown in Figure 1. Alternative mappings have been considered [2], but are unlikely to be adopted because some dialing systems require adherence to the ITU standard. Therefore, we consider only the ITU keypad mappings here.

The T9 mapping associates a single key with multiple letters, introducing ambiguity in key-entry, which must be resolved by the input interface. Early systems employed *multi-tap* interfaces, where each letter is unambiguously identified by a number of successive "taps" on a numeric key. More recently, multi-tap interfaces have been superceded by *single-tap* – or "predictive text" – interfaces[1], which have been shown to reduce the number of keystrokes required per character [3].

[1] In many single-tap interfaces, multi-tap entry is still used to enter novel words.

Advances in Natural Language Processing and Applications
Research in Computing Science 33, 2008, pp. 237-248

Received 31/10/07
Accepted 07/12/07
Final Version 17/01/08