Scaling to Billion-plus Word Corpora

Jan Pomikálek[1], Pavel Rychlý[1] and Adam Kilgarriff[2]

[1] Masaryk University, Brno, Czech Republic
[2] Lexical Computing Ltd, Brighton, UK

Abstract. Most phenomena in natural languages are distributed in accordance with Zipf's law, so many words, phrases and other items occur rarely and we need very large corpora to provide evidence about them. Previous work shows that it is possible to create very large (multi-billion word) corpora from the web. The usability of such corpora is often limited by duplicate contents and a lack of efficient query tools.
This paper describes BiWeC, a Big Web Corpus of English texts currently comprising 5.5b words fully processed, and with a target size of 20b. We present a method for detecting near-duplicate text documents in multi-billion-word text collections and describe how one corpus query tool, the Sketch Engine, has been re-engineered to efficiently encode, process and query such corpora on low-cost hardware.

1 Introduction

There's no data like more data, and one place to get more data almost without limit (for general English and some other languages and varieties) is the web. One way to use the web is to create a local corpus by downloading web pages: in [1] we argue that it is the optimal way to use the web for linguistic research. A number of corpora have been built in this way: Baroni and colleagues developed web corpora with nearly 2 billion words for German, Italian and English [2, 3] and have made them available for research as tar archives. Liu at al. [4] describe the creation of a 10 billion word corpus. In this paper we introduce BiWeC, a Big Web Corpus currently of 5.5b words, with a target size of 20b.

Very large corpora can be created on low cost hardware in a few person-months. Most of the steps have linear complexity and scale up well. The two outstanding issues we focus on in this paper are:

1. removing duplicate content
2. efficient querying.

The article is organized as follows. In section two our motivation for creating larger corpora is discussed and the advantages of using more data for various tasks is explained. Section three describes the process of creating BiWeC, with a focus on removing duplicate and near-duplicate documents. Section four deals with corpus processing and querying using the Sketch Engine corpus manager. Then we present some figures about BiWeC and outline future plans.

Detecting and grounding terms in biomedical literature

Kaarel Kaljurand, Fabio Rinaldi, Thomas Kappeler, Gerold Schneider

Institute of Computational Linguistics, University of Zurich
kalju@ifi.uzh.ch, rinaldi@ifi.uzh.ch, kappeler@bluewin.ch,
gschneid@ifi.uzh.ch

Abstract. We present an approach towards the automatic detection of names of proteins, genes, species, etc. in biomedical literature and their grounding to widely accepted identifiers. The annotation is based on a large term list that contains the common expression of the terms, a normalization step that matches the terms with their actual representation in the texts, and a disambiguation step that resolves the ambiguity of matched terms. We describe various characteristics of the terms found in existing term resources and of the terms that are used in biomedical texts. We evaluate our results against a corpus of manually annotated protein mentions and achieve a precision of 57% and recall of 72%.

1 Introduction

The complexity of biological organisms and the success of biological research in describing them, have resulted in a large body of biological entities (genes, proteins, species, etc.) to be indexed, named and analyzed. Probably the most important entities are proteins. They are an essential part of an organism and participate in every process within cells. Most proteins function in collaboration with other proteins, and one of the research goals in molecular biology is to identify which proteins interact.

While the number of different proteins is large, the amount of their possible interactions and combinations is even larger. In order to record such interactions and represent them in a structured way, human curators who work for knowledge base projects, e.g. Molecular INTeraction database (MINT)[1], Human Protein Reference Database (HPRD)[2], IntAct[3] (see [4] for a detailed overview), carefully analyze published biomedical articles. As the body of articles is growing rapidly, there is a need for effective automatic tools to help curators in their work. Such tools must be able to detect mentions of biological entities in the text and tag them with identifiers that have been assigned by existing knowledge bases. As the names that are used to reference the proteins can be very ambiguous, there is a need for an effective ambiguity resolution.

[1] http://mint.bio.uniroma2.it
[2] http://www.hprd.org/
[3] http://www.ebi.ac.uk/intact

Automatic Word Clustering in Studying Semantic Structure of Texts

Olga Mitrofanova [1]

[1] St. Petersburg State University, Faculty of Philology and Arts,
Department of Mathematical Linguistics,
Universitetskaya emb., 11
199034 St. Petersburg, Russia
{alkonost-om@yandex.ru}

Abstract. The purpose of the study is to prove that results of automatic word clustering (AWC) may contribute much in investigating semantic structure of texts and in evaluating plot complexity. Experiments were carried out for Russian texts, mainly stories and short novels. Data obtained in course of study allowed to formulate and verify several linguistic hypotheses.

Keywords: Automatic Word Clustering, Russian Corpora, Semantic Structure of Texts

1 Introduction

Formalization of text structure and quantitative evaluation of semantic relations between text units prove to be of considerable importance in various fields of natural language understanding: modelling plot structure, text summarization, evaluation of translation adequacy in parallel texts, automatic text indexing, classification of texts in corpora, etc. (for a detailed analysis cf. [1], [2]).

One of the procedures providing linguistic data on semantic structure of texts is automatic word clustering (AWC). It is assumed that AWC results help to reveal semantic structure of texts and to determine plot complexity. To prove this assumption, AWC procedure was carried out with the help of a specialized AWC toolkit based on word space model. Experimental procedure implied processing Russian texts, mainly stories and short novels. A set of key words describing major topics of the plot was assigned to each text, clusters of words with similar distributions were created for each key word. Data extracted from texts through AWC procedure admit thorough linguistic interpretation. Further comparison of cluster content and structure allowed to distinguish texts characterized by a plot including a dominating topic with a number of subtopics and texts characterized by a plot including a set of major (independent or correlating) topics.

Semantically-Driven Extraction of Relations between Named Entities

Caroline Brun and Caroline Hagège

Xerrox Research Centre Europe, 6 chemin de Maupertuis
38240 Meylan, France
{Caroline.Brun, Caroline.Hagege}@xrce.xerox.com

Abstract. In this paper, we describe a method that automatically generates lexico-syntactic patterns which are then used to extract semantic relations between named entities. The method uses a small set of seeds, i.e. named entities that are a priori known to be in relation. This information can easily be extracted from encyclopedias or existing databases. From very large corpora we extract sentences that contain combinations of these attested entities. These sentences are then used in order to automatically generate, using a syntactic parser, lexico-syntactic patterns that links these entities. These patterns are then re-applied on texts in order to extract relations between new entities of the same type. Furthermore, the patterns that are extracted not only provide a way to spot new entities relations but also build a valuable paraphrase resource. An evaluation on the relation holding between an event, the place of the event occurrence and the date of the event occurrence has been carried out on French corpus and shows good results. We believe that this kind of methodology can be applied for other kinds of relation between named entities.

1 Introduction

In this paper we describe a system that extracts accurately semantic relations between named entities from raw text. Taking as input a small set of already known relations that can be extracted from encyclopedias or from databases, our system first learn from a large corpus a wide range of lexico-syntactic patterns conveying the desired semantic relation. These learned patterns are then further applied on texts, and as a result, new occurrences of the given semantic relations linking new entities are detected. As all the patterns extracted represent a comparable semantic situation, they can be considered as paraphrase patterns. These patterns can then be used both in generation and for information extraction tasks.

2 Related Work

Many research works on extraction of relation between entities have already been performed since this kind of information is useful for a wide range of applications of information extraction. For instance [8] describe an algorithm to extract relations between named entities and the resulting improvement of a question answering system. Semantic relation detection between named entities has also been investigated in the context of the semantic web (try to obtain a rich and accurate metadata annotation from web content) as for instance in [6]. In the biomedical domain, [7] and [14] present methods to automatically extract interaction relations between genes and/or protein using machine learning techniques.

Some of these approaches rely on pattern matching exploiting simple syntactic relations as the Subject-Verb-Object relation. Sometimes, additional ontological knowledge is also exploited. These approaches take advantage of the fact that a certain syntactic configuration can be mapped onto a semantic relation. This is particularly well described in [12] where a shallow parser and a deeper parser are used to extract SVO relations between triples. In [7] and [14], a previous dependency analysis is performed to derive necessary information for learning algorithms.

Evaluation of Named Entity Extraction Systems

Mónica Marrero, Sonia Sánchez-Cuadrado, Jorge Morato Lara, George Andreadakis

Computer Engineering Department, University Carlos III of Madrid
Av. de la Universidad 30, 28911 Leganés (Madrid), Spain
mmarrero@inf.uc3m.es, ssanchec@ie.inf.uc3m.es, jmorato@inf.uc3m.es, gand@ie.inf.uc3m.es

Abstract. The suitability of the algorithms for recognition and classification of entities (NERC) is evaluated through competitions such as MUC, CONLL or ACE. In general, these competitions are limited to the recognition of predefined entity types in certain languages. In addition, the evaluation of free applications and commercial systems that do not attend the competitions has been lightly studied. Shallowly studied have also been the causes of erroneous results. In this study a set of NERC tools are assessed. The assessment of the tools has consisted of: 1) the elaboration of a test corpus with typical and marginal types of entities; 2) the elaboration of a brief technical specification for the tools evaluated; 3) the assessment of the quality of the tools for the developed corpus by means of precision-recall ratios; 4) the analysis of the most frequent errors. The sufficiency of the technical characteristics of the tools and their evaluation ratios, presents an objective perspective of the quality and the effectiveness of the recognition and classification techniques of each tool. Thus, the study complements the information provided by other competitions and aids the choice or the design of more suitable NER tools for a specific project.

Keywords: Named entity extraction, named entity recognition and classification, information extraction, named entity extraction tools.

Introduction

There is currently a wide variety of named entities (NE) recognition systems. Competitive events are organized for the evaluation of NERC systems, in which the ability of identification and classification of the entities existing in a corpus is analyzed. Nevertheless, the competitions normally establish certain limitations such as:

– They focus on a limited group of NE types. This feature is quite variable due to the ambiguity in the use of the term *Named Entity* depending on the different forums or events. In the case of the MUC conferences, NEs were considered *personal names*, *organizations*, *locations* and at a later stage, *temporal entities* and *measurements* [1]. On the other hand, the CONLL-2002/2003 conferences defined the categories *person*, *organism*, *localization* and *miscellaneous* [2, 3]. The latter (*miscellaneous*), includes proper names of different nature with different categories*: gentilics*, *project names*, *team names*, etc. Finally, the ACE

Building a Brazilian Portuguese parallel corpus of original and simplified texts

Helena M. Caseli[1], Tiago F. Pereira[1], Lucia Specia[1], Thiago A. S. Pardo[1], Caroline Gasperin[1], and Sandra M. Aluisio[1]

[1]Center of Computational Linguistics (NILC)/ Department of Computer Sciences, University of São Paulo, Av. Trabalhador São-Carlense, 400. 13560-970 - São Carlos/SP, Brazil
helenacaseli@dc.ufscar.br, tiagofrepereira@yahoo.com.br, lspecia@icmc.usp.br, taspardo@icmc.usp.br, cgasperin@icmc.usp.br, sandra@icmc.usp.br

Abstract. In this paper we address the problem of building the necessary tools and resources for performing Brazilian Portuguese text simplification. We describe our efforts on the design and development of: (a) a XCES-based annotation schema, (b) an annotation edition tool, and (c) a portal to access parallel corpora of original-simplified texts. These contributions were intended to (i) allow the creation and public release of a corpus of original and simplified texts with two different versions of simplification (called here *natural* and *strong*), targeting two levels of functional illiteracy and (ii) register simplification decisions during the creation of such corpus. We also provide an analysis of the first corpus created using the resources presented here: 104 newspaper texts and their simplified versions, produced by an expert in text simplification.

Keywords: Text Simplification, Brazilian Portuguese, annotation standards, annotation edition tool.

1 Introduction

In Brazil, "letramento" (literacy) is the term used to designate people's ability to use written language to obtain and record information, express themselves, plan and learn continuously [1]. In Brazil, according to the index used to measure the literacy level of the population (*INAF - National Indicator of Functional Literacy*), a vast number of people belong to the so called *rudimentary* and *basic* literacy levels. These people are able to find explicit information in short texts (rudimentary level) and also process slightly longer texts and make simple inferences (basic level).

The PorSimples project (*Simplificação Textual do Português para Inclusão e Acessibilidade Digital*)[1] aims at producing text simplification tools for promoting digital inclusion and accessibility for people with such levels of literacy, and possibly other kinds of reading disabilities. More specifically, the goal is to help these readers

[1] http://caravelas.icmc.usp.br/wiki/index.php/Principal

Synchronized Morphological and Syntactic Disambiguation for Arabic

Daoud Daoud

Pricess Sumaya University for Technology
Daoud@batelco.jo

Abstract. In this paper, we present a unique approach to disambiguation Arabic using a synchronized rule-based model. This approach helps in highly accurate analysis of sentences. The analysis produces a semantic net like structure expressed by means of Universal Networking Language (UNL)- a recently proposed interlingua. Extremely varied and complex phenomena of Arabic language have been addressed.

Keywords: Arabic Language, Synchronized Model, Disambiguation, UNL

1. INTRODUCTION

Compared to French or English, Arabic as an agglutinative and highly inflected language shows its proper types of difficulties in morphological disambiguation, since a large number of its ambiguities come from both the stemming and the categorization of a morpheme while most of ambiguities in French or English are related to the categorization of a morpheme only.
Phrases and sentences in Arabic have a relatively free word. The same grammatical relations can have different syntactic structures. Thus, morphological information is crucial in providing signs for structural dependencies.
Arabic sentences are characterized by a strong tendency for agreement between its constituents, between verb and noun, noun and objective, in matters of numbers, gender, definitiveness, case, person etc. These properties are expressed by a comprehensive system of affixation.
Arabic uses a diverse system of prefixes, suffixes, and pronouns that are attached to the words, creating compound forms that further complicate text manipulation. Simultaneously, Arabic exhibits a large-scale ambiguity already at the word level, which means that there are multiple ways in which a word can be categorized or broken down to its constituent morphemes. This is further complicated by the fact that most vocalization marks (diacritics) are omitted in Arabic texts.
However, the morphological analysis of a word-form, and in particular its morphological segmentation, cannot be disambiguated without reference to context,

Parsing with Polymorphic Categorial Grammars

Matteo Capelletti[1] and Fabio Tamburini[2]

[1] Lix, École Polytechnique - France
matteo.capelletti@elisanet.fi
[2] DSLO, University of Bologna - Italy
fabio.tamburini@unibo.it

Abstract. In this paper we investigate the use of polymorphic categorial grammars as a model for parsing natural language. We will show that, despite the undecidability of the general model, a subclass of polymorphic categorial grammars, which we call *linear*, is mildly context-sensitive and we propose a polynomial parsing algorithm for these grammars.

1 Introduction

The simplest model of a categorial grammar is the so called Ajdukiewicz–Bar-Hillel calculus of [2] and [4]. Syntactic categories are formed from a given set of atoms as *functions* a/b and $b\backslash a$, with b and a categories. The intuitive meaning of a syntactic category of the form a/b (resp. $b\backslash a$) is that it looks for an *argument* of category b to its right (resp. left) to give a category of type a. The resulting grammar system is known to be context-free.

Contemporary categorial grammars in the style of Ajdukiewicz–Bar-Hillel grammars are called *combinatory categorial grammars*, see [25]. Such systems adopt other forms of composition rules which enable them to generate non–context-free languages, see [29; 28]. The other main tradition of categorial grammar, the type-logical grammars of [20; 18], stemming from the work of [15], adopt special kinds of structural rules, that enable the system to generate non–context-free languages.

Both approaches increase the generative power of the basic system by adding special kinds of rules. In this paper, we adopt a different strategy which consists in keeping the elementary rule component of Ajdukiewicz–Bar-Hillel grammar and in introducing *polymorphic* categories, that is syntactic categories that contain variables ranging over categories. The inference process will be driven by unification, rather than by simple identity of formulas. We will see two kinds of polymorphic categorial grammars, one that is Turing complete and another, resulting from a restriction on the first, which is mildly context-sensitive. This second system, which is obviously the most interesting one for linguistics, has some important advantages with respect to the aforementioned ones. In respect to TLG, the polymorphic system we define is *polynomial*, as we will prove by providing a parsing algorithm. In respect to CCG (and most known TLG), our system is not affected by the so called *spurious ambiguity* problem, that is the problem of generating multiple, semantically equivalent, derivations.

A CCG-based System for Valence Shifting for Sentiment Analysis

František Simančík and Mark Lee
School of Computer Science
University of Birmingham
Birmingham, UK
frantisek.simancik@worc.ox.ac.uk, m.g.lee@cs.bham.ac.uk

Abstract

The automatic classification of sentiment in text is becoming an important area of research. In this work, we present a linguistic system for sentence-level valence annotation. Our system uses the formalism of Combinatory Categorial Grammar to represent words as functions acting on their syntactic arguments, which provides a unified way of implementing various classes of valence shifters. We propose two simple semi-automatic methods for estimating the valence of individual terms based on the lexical relations of WordNet. We evaluate the system on the data generated for the Affective Text task of SemEval 2007 and show that it compares favourably with the systems participating in the task.

Keywords: sentiment analysis, valence annotation, valence shifters, headlines, combinatory categorial grammar.

1 Introduction

The number of opinion-rich resources such as discussions, blogs and review sites has been growing rapidly in recent years. As a result of this, there is a demand for tools capable of classifying texts not only by the topic but also the attitude and opinion they convey; giving rise to new areas in Natural Language Processing called Opinion Mining and Sentiment Analysis.

One of the most prominent tasks in the field is the classification of valence (positive/negative orientation). Researchers (Pang et al. [7], Kennedy and Inkpen [6] and others) have successfully applied supervised machine learning methods[1] to determine the valence of longer texts. These approaches rely on the availability of a large amount of human-tagged training data and, compared to linguistic methods, reveal very little about the nature of the connection between a text and the opinion it expresses.

[1] Naive Bayes Classifiers, Support Vector Machines, etc.

Classification of "Inheritance" Relations: a Semi-automatic Approach

Ekaterina Lapshinova-Koltunski

IMS, Universität Stuttgart
Azenbergstr.12
70174 Stuttgart
katerina@ims.uni-stuttgart.de

Abstract. This study describes a semi-automatic approach to the classification of "inheritance" relations between morphologically related predicates.
Predicates, such as verbs and nouns subcategorizing for a subclause, are automatically extracted from text corpora and are classified accroding to their subcategorisation properties. For this purpose, we elaborate a semi-automatic knowledge-rich extraction and classification architecture. Our aim is also to compare subcategorisation properties of morphologically related predicates, i.e. verbs and deverbal nouns.
In this work, we concentrate exclusively on the predicates with sentential complements, such as *dass, ob* and *w*-clauses (that, if and wh-clauses) in German, although our methods can be applied for other complement types as well.

1 Introduction

This paper describes a semi-automatic approach to the analysis of subcategorisation properties of morphologically related predicates, such as verbs and nouns. We classify predicates according to their subcategorisation properties by means of extracting them from German corpora along with their complements. In this work, we concentrate exclusively on sentential complements, such as *dass, ob* and *w*-clauses, although our methods can be also applied for other types of complements.

It is usually assumpted that subcategorisation properties of nominalisations are taken over from their underlying verbs. However, our preliminary tests show that there exist different types of relations between them. Thus, our aim is to review the properties of morphologically related words and to analyse the phenomenon of "inheritance" of subcategorisation properties.

For this purpose, we elaborate a set of semi-automatic procedures, with the help of which we not only classify extracted units according to their subcategorisation properties, but also compare the properties of verbs and their nominalisations. Our aim is to serve NLP, especially such large symbolic grammar for deep processing as HPSG or LFG, which need detailed subcategorisation data for their lexicons and grammars.

Discovering Discourse Motifs in Instructional Dialog

Juan M. Huerta[1]

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

Abstract. We propose a method to analyze conversational interaction using discourse motifs (sequence of labels). We focus specifically on instructional transactive discourse. We first describe the characteristics of transactive discourse, its relationship to other frameworks of instructional discourse, and introduce a refined taxonomy of transactive discourse. Based on this new taxonomy, we construct a set of classifiers to automatically label instructional dialog segments. After labeling, we search for salient patterns of discourse common to these chains of labels using Multiple EM for Motif Elicitation and Gapped Local Analysis of Motifs (which are two techniques available for DNA and protein motif discovery). From our analysis of a corpus of classroom data, a set of Transactive-Participatory-Coherent motifs emerge. This approach to interaction-motif discovery and analysis can find application in dialog and discourse analysis, pedagogical domains (e.g., assessment and professional development), automatic tutoring systems, meeting analysis, problem solving, etc.

1 Introduction

We focus on the analysis of classroom discourse particularly when the focus is on solving mathematical problems. While the analysis of classroom discourse and mathematical problem solving is useful in providing pedagogical insight into teaching practices (see for example Huerta (2008), Blanton (2008)), its analysis can also shed light into interaction mechanisms used in more general collaborative problem solving.

Research in human dialog has been approached from various viewpoints using frameworks and methodologies of analysis that have been tailored to address the specific requirements of these viewpoints (examples of relatively recent perspectives to dialog analysis include Stolcke (2000), Stent (2000) among others, and a good summary can be found in Moore (2003)).

More problem-solving specific frameworks have also been proposed to analyze planning-oriented and instructional dialog in the classroom (Linden (1995)). Additionally, there have also been other efforts in the manual analysis of classroom interaction from purely pedagogical and sociological perspectives (Blanton (2008), Mehan (1985), Stark (2002), Haussman (2003)). There has been also work focusing on specific theoretical frameworks of interaction and the correlation of their elements to

Ontology Oriented Computation of English Verbs Metaphorical Trait

Zili Chen, Jonathan J. Webster, Ian I. Chow, Tianyong Hao

Department of Chinese, Translation and Linguistics, City University of Hong Kong,
81 Tat Chee Avenue, Kowloon Tong, KLN, Hong Kong

Abstract. Research on metaphor has generally focused on exploring its context-dependent behavior and function. This current study aims to testify the postulate of English verb's innate trait of Metaphor Making potential. This paper intends to carry out an in-depth case study of a group of English core verbs using WordNet and SUMO ontology. In order to operationalize the assessment of an English verb's metaphor making potential, a refined algorithm has been developed, and a program made to realize the computation. At last, it is observed that higher frequency verbs generally possess greater metaphor making potential; while a verb's metaphor making potential on the other hand is also strongly influenced by its functional categories. As a preliminary context-free experiment with metaphor, this research foresees the possibility of providing an annotation schema for critical discourse analysis and a new parameter for scaling the difficulty level of reading comprehension of English texts.

Keywords: ontological computation, English verbs, MMP

1 Introduction and Previous Work

Metaphorical computation continues to remain a significant challenge to NLP. Recent researches of it mainly fall into two categories: rule-based approaches and statistical-based approaches. Up to now, some achievements have been attained, among which knowledge representation based methods are predominant [1]. These methods mainly employ knowledge representation based ontologies, such as The Suggested Upper Merged Ontology (SUMO), as their working mechanism. However, those researches are all limited to the study of metaphor's behavior and function in different contexts.

In line with Lakoff's view [2], "Metaphor allows us to understand one domain of experience in terms of another. This suggests that understanding takes place in terms of entire domains of experience and not in terms of isolated concepts", SUMO, an effort of the IEEE Standard Upper Ontology Working Group with the support of Teknowledge, contains terms chosen to cover all general domain concepts needed to represent world knowledge. Whereas Ahrens & Huang's research with SUMO and metaphor has focused on specific domain metaphors [3, 4], thus failing to make full use of SUMO's overall domain coverage.

Automatic formal verification of conceptual model documentation by means of self-organizing map

Algirdas Laukaitis and Olegas Vasilecas

Vilnius Gediminas Technical University , Sauletekio al. 11,
LT-10223 Vilnius-40, Lithuania
{algirdas.laukaitis,olegas}@fm.vgtu.lt

Abstract. By using background knowledge of the general and specific domains and by processing new natural language corpus experts are able to produce a conceptual model for some specific domain. In this paper we present a model that tries to capture some aspects of this conceptual modeling process. This model is functionally organized into two information processing streams: one reflects the process of formal concept lattice generation from domain conceptual model, and the another one reflects the process of formal concept lattice generation from the domain documentation. It is expected that similarity between those concept lattices reflects similarity between documentation and conceptual model.In addition to this process of documentation formal verification the set of natural language processing artifacts are created. Those artifacts then can be used for the development of information systems natural language interfaces. To demonstrate it, an experiment for the concepts identification form natural language queries is provided at the end of this paper.

Key words: Information systems engineering, formal concept analysis, IS documents self-organization, natural language processing.

1 Introduction

Software engineers spend hours in defining information systems (IS) requirements and finding common ground of understanding. The overwhelming majority of IS requirements are written in natural language supplemented with conceptual model and other semi-formal UML diagrams. The bridge in the form of semantic indexes between documents and conceptual model can be useful for more effective communication and model management. Then, an integration of the natural language processing (NLP) into information system documentation process is an important factor in meeting challenges for methods of modern software engineering.

Reusing natural language IS requirement specifications and compiling them into formal statements has been an old challenge [1], [14]. Kevin Ryan claimed that NLP is not mature enough to be used in requirements engineering [13] and

Coreference resolution using Markov Logic Networks

Shujian HUANG[1], Yabing ZHANG[1], Junsheng ZHOU[12], and Jiajun CHEN[1]

[1] State Key Laboratory for Novel Software Technology,
Nanjing University, Nanjing 210093, China
[2] Department of Computer Science, Nanjing Normal University,
Nanjing, Jiangsu, 210097, China
{huangsj, zhangyb, zhoujs, chenjj}@nlp.nju.edu.cn

Abstract. Most previous work treats the solution for pronouns and noun phrases either in two separate processes or in a single process. We argue that resolving them in two processes may result in the loss of potential useful information for each process. However, resolving them in a single process is also problematic. These two types of mentions have very different characteristics in some commonly used features. Current models cannot catch those differences and thus the two types may interfere with each other. In this paper, we propose a modeling strategy using Markov logic networks (MLNs) which can explicitly discriminate the two types in one single process. Experiments on ACE2005 Chinese dataset show that our modeling using MLNs, together with the correlation clustering technique, brings significant improvements to the task.

Key words: Coreference Resolution, Markov Logic Networks

1 Introduction

Coreference resolution (CR) has drawn a lot of attentions over the past decade, especially since McCarthy[1], Cardie and Wagstaff[2] introduced machine learning techniques into this field. It plays an important role in understanding complex texts and is widely used in a lot of applications such as question answering[3], summarization[4], etc. The strong relation with other popular topics such as entity resolution in database and citation analysis[5] makes it more attractive.

Pronoun and noun phrase are two major types of mentions in CR. There are two strategies for the resolution of them. One strategy tends to split the resolution of pronouns and noun phrases into two separate processes (*Separate Strategy*). Some works focus on just pronoun resolution, aiming to find the right antecedent for each pronoun[6–9]. Denis and Baldridge subdivide mentions into five categories such as third person pronouns, speech pronouns, etc. Then, specialized models are proposed for each individual type[10]. However, we argue that just considering pairwise relation between pronoun and each of its antecedent candidates does not make full use of the information among those candidate

A Ranking Approach to Persian Pronoun Resolution

Nafiseh Sadat Moosavi and Gholamreza Ghassem-Sani

Department of Computer Engineering, Sharif University of Technology, Iran
n_moosavi@ce.sharif.edu, sani@sharif.edu

Abstract. Coreference resolution is an essential step toward understanding discourses, and it is needed by many NLP tasks such as machine translation, question answering, and summarization. Pronoun resolution is a major and challenging subpart of coreference resolution, in which only the resolution of pronouns is considered. Classification approaches have been widely used for coreference/pronoun resolution, but it has been shown that ranking approaches outperform classification approaches in a variety of fields such as English pronoun resolution (Denis and Baldridge, 2007), question answering (Ravichandran, 2003), and tagging/parsing (Collins and Duffy, 2002; Charniak and Johnson, 2005). The strength of ranking is in its ability to consider all candidates at once and selecting the best one based on the model, while existing classification methods consider at most two candidate responses at a time. Persian and its varieties are spoken by more than 71 million people, and it has some characteristic that make parsing and other related processing of Persian more difficult than those of English. In this paper, we have evaluated maximum entropy ranker on Persian pronoun resolution and compared the results with that of four base classifiers.

Keywords: Natural Language Processing, Machine Learning, Ranking, Classification, Pronoun Resolution, Persian.

1 Introduction

The final goal of natural language processing (NLP) is that computers understand human languages. Different NLP research areas such as part of speech (POS) tagging, word sense disambiguation (WSD), and grammatical parsing concentrate only on a partial solution of this ultimate goal. All of these are required for a computer to understand a natural language.

NLP tasks can be divided into micro-tasks and macro-tasks. Micro-tasks focus on a word level processing or a sentence level processing such as WSD and parsing. On the other hand, macro-tasks include tasks which do a document level processing such as information retrieval and document classification. Before the introduction of machine learning approaches in NLP, higher level tasks such as semantic processing needed a variety of lower level tasks such as POS tagging and parsing. However, the use of machine learning methods may make it possible to obtain enough statistical in-

Classification of Clinical Conditions: A Case Study on Prediction of Obesity and Its Co-morbidities

Archana Bhattarai, Vasile Rus, Dipankar Dasgupta

Department of Computer Science, The University of Memphis,
209 Dunn Hall
Memphis, TN 38152-3240, USA
{abhattar, vrus, dasgupta}@memphis.edu

Abstract. We investigate a multiclass, multilabel classification problem in medical domain in the context of prediction of obesity and its co-morbidities. Challenges of the problem not only lie in the issues of statistical learning such as high dimensionality, interdependence between multiple classes but also in the characteristics of the data itself. In particular, narrative medical reports are predominantly written in free text natural language which confronts the problem of predominant synonymy, hyponymy, negation and temporality. Our work explores the comparative evaluation of both traditional statistical learning based approach and information extraction based approach for the development of predictive computational models. In addition, we propose a scalable framework which combines both the statistical and extraction based methods with appropriate feature representation/selection strategy. The framework leads to reliable results in making correct classification. The framework was designed to participate in the second i2b2 Obesity Challenge.

Keywords: Text classification, Information Extraction, natural language processing

1 Introduction

Medical informatics is chiefly inductive and information intensive science where observation and analysis of comprehensive clinical data can lead to complex and powerful evidence based decision support systems. One of the primary goals of these automated systems is to make information more accessible, representative and meticulous in a quick span[4]. Furthermore, they have gained increased importance in the recent years as it can even outperform a human expert in some cases in diagnosing diseases as the process is highly subjective and fundamentally depends on the experiences of the assessor and his/her interpretation on the information[4]. Conversely, most medical institutions are still keeping a large amount of medical data in narrative form resulting in huge volume of potential information with limited or no utility and accessibility. An effort to exploit this data poses multiple challenges as it involves processing free text data with the presence of acronyms, synonyms, negation

Analysis of Stemming Alternatives and Dependency Pattern Support in Text Classification

Levent Özgür and Tunga Güngör

Department of Computer Engineering, Boğaziçi University,
Bebek, 34342 Istanbul,Turkey
{ozgurlev,gungort}@boun.edu.tr

Abstract. In this paper, we study text classification algorithms by utilizing two concepts from Information Extraction discipline; dependency patterns and stemmer analysis. To the best of our knowledge, this is the first study to fully explore all possible dependency patterns during the formation of the solution vector in the Text Categorization problem. The benchmark of the classical approach in text classification is improved by the proposed method of pattern utilization. The test results show that support of four patterns achieves the highest ranks, namely, *participle modifier*, *adverbal clause modifier*, *conjunctive* and *possession modifier*. For the stemming process, we benefit from both morphological and syntactic stemming tools, Porter stemmer and Stanford Stemmer, respectively. One of the main contributions of this paper is its approach in stemmer utilization. Stemming is performed not only for the words but also for all the extracted pattern couples in the texts. Porter stemming is observed to be the optimal stemmer for all words while the raw form without stemming slightly outperforms the other approaches in pattern stemming. For the implementation of our algorithm, two formal datasets, Reuters - 21578 and National Science Foundation Abstracts, are used.

Key words: Text Classification, Dependency Patterns, Stemmer Analysis, Information Extraction

1 Introduction

Text Classification (TC) is a learning task, where pre-defined category labels are assigned to documents based on the likelihood suggested by a training set of labelled documents.

Most of the approaches used in this problem study it in bag-of-words (bow) form, where only the words in the text are analyzed by some machine learning algorithms for TC [1]. In this approach, documents are represented by the widely used vector-space model, introduced by Salton et al. [2]. In this model, each document is represented as a vector d. Each dimension in the vector d stands for a distinct term (word) in the term space of the document collection.

Investigating Variations in Adjective Use across Different Text Categories

Jing Cao[1] and Alex Chengyu Fang[2]

Department of Chinese, Translation and Linguistics
City University of Hong Kong
Hong Kong SAR, China
[1]cjing3@student.cityu.edu.hk, [2]acfang@cityu.edu.hk

Abstract. Adjectives are an informative but understudied linguistic entity with good potentials in sentiment analysis, text classification and automatic genre detection. In this article, we report an investigation of the variations in adjective use across different text categories represented in a sizable corpus. In particular, we report the distribution of adjectives across a range of categories grouped together as academic propose in the British National Corpus. We shall measure inter-category similarity in the use of adjectives and demonstrate with empirical data that adjectives are an effective differentia of text categories or domains, at least in terms of arts and sciences as the two major sub-categories within academic prose.

Key Words: corpus, text category, adjective, similarity, BNC

1 Introduction

Adjectives are an informative but understudied linguistic entity [1, 2], drawing more and more attention within the research community. Focus has been mostly on the semantic aspect of adjectives for practical research in sentiment analysis applicable to automatic evaluations of email communication [3], blogs [4] and customer reviews in [5]. Studies in this respect typically focus on evaluative adjectives [6] and size adjectives [7]. In addition to the semantic approach, adjectives are also used for purposes of text categorization and genre detection in [8]. In this respect, [2] and [9] have generally shown with corpus evidence that adjectives occur more often in written texts than in spoken ones, and more frequently in informative writing than in imaginative writing. According to [8], 'the literature suggests that adjectives and adverbs will vary by genre because of their unique patterns of usage in text' (p. 4).

This paper describes one of the recent attempts to study adjectives from the perspectives of text categorization and genre detection. In particular, we investigate the variations of adjective use across various types of academic writing selected from a large-sized corpus. We attempt to ascertain whether adjective-based indices will be able to classify texts in such a way that conforms to manual classification. As we shall show in this article with empirical data, adjectives do differ by text categories and

Multi-Category Support Vector Machines for Identifying Arabic Topics

Mourad Abbas, Kamel Smaili and Daoud Berkani

CRSTDLA, Speech Processing Lab.
1 rue D. E. Alafghani, Algeria
INRIA-LORIA, Parole team
B.P. 101, Villers les Nancy, France
Polytechnic School, Signal and Communication Lab.
10 rue H. Badi, Algeria
m_abbas04@yahoo.fr, kamel.smaili@loria.fr, dberkani@enp.edu.dz

Abstract. It is known that Support Vector Machines were designed for binary classification. Nevertheless, it would be fruitful to extend this operation to what is called Multi-category classification. That is why Multi-category Support Vector Machines (MSVM) become nowadays the current subject of several serious researches, aiming to achieve high levels of multi-category classification tasks. This technique has been assessed recently in some fields as text categorization, Cancer classification, etc. We should notify that experiments which have been realized until now using MSVM are limited to small data sets, since its computation is more expensive. In this paper we are interested in the use of this method, for the first time in topic identification. The experiments conducted concern topic identification of Arabic language. The corpora are extracted from Alwatan newspaper. Achieved results lead to an improvement of MSVM performance in comparison to the baseline SVM method. Nevertheless, SVM still outperforms MSVM when using larger sizes of the vocabulary.

1 Introduction

The main objective of topic identification is to assign one or several topic labels to a flow of textual data. Labels are chosen from a set of topics fixed a priori. Talking about topics conduct us to clarify the definition of a topic. In [1], each keyword is considered as a topic. Whereas in other works, topics are more sophisticated corresponding to specific subject, for example politics and sports [2]. In our case, we are dealing with six topics: Culture, Religion, Economy, Local news, International news and sports.
Topic identification is used in several areas: to adapt language models for speech recognition and for machine translation, to focus on a specific use for search engines,...etc. In spontaneous speech recognition process the vocabulary has to be as large as possible. Enlarging the vocabulary increases the search space and consequently could reduce system's performance.
A language model is one of the knowledge sources which is used by a speech

USUM: Update Summary Generation System

C Ravindranath Chowdary and P Sreenivasa Kumar

Department of Computer Science and Engineering
Indian Institute of Technology Madras
Chennai 600 036, India
{chowdary,psk}@cse.iitm.ac.in

Abstract. *Huge amount of information is present in the World Wide Web and a large amount is being added to it frequently. A query-specific summary of multiple documents is very helpful to the user in this context. Currently, few systems have been proposed for query-specific, extractive multi-document summarization. If a summary is available for a set of documents on a given query and if a new document is added to the corpus, generating an updated summary from the scratch is time consuming and many a times it is not practical/possible. In this paper we propose a solution to this problem. This is especially useful in a scenario where the source documents are not accessible. We cleverly embed the sentences of the current summary into the new document and then perform query-specific summary generation on that document. Our experimental results show that the performance of the proposed approach is good in terms of both quality and efficiency.*

1 Introduction

Currently, the World Wide Web is the largest source of information. Huge amount of data is present on the Web and large amount of data is added to the web constantly. Often the information pertaining to a topic is present across several web pages. It is a tedious task for the user to go through all these documents as the number of documents available on a topic will range from tens to thousands. It will be of great help for the user if a query specific multi-document summary is generated. Summary generation can be broadly divided as abstractive and extractive. In abstractive summary generation, the abstract of the document is generated. The summary so formed need not have exact sentences as present in the document. In extractive summary generation, important sentences are extracted from the document. The generated summary contains all such extracted sentences arranged in a meaningful order. In this paper, generated summaries are extractive. Summary can be generated either on a single document or on several documents. In multi-document summary generation, other issues like time, ordering of extracted sentences, scalability etc. will arise.

Summary can be either generic or query specific. In generic summary generation, the important sentences from the document are extracted and the sentences

Towards the Speech Synthesis of Raramuri: a unit selection approach based on unsupervised extraction of suffix sequences

Alfonso Medina Urrea,[1] José Abel Herrera Camacho[2], and Maribel Alvarado García[3]

[1] GIL-II, Universidad Nacional Autónoma de México
04510 Coyoacán, DF, MEXICO
amedinau@ii.unam.mx
[2] FI, Universidad Nacional Autónoma de México
04510 Coyoacán, DF, MEXICO
abelh@verona.fi-p.unam.mx
[3] Escuela Nacional de Antropología e Historia
14030 Tlalpan, DF, MEXICO
marvarado1978@yahoo.com.mx

Abstract. This work deals with the design of a synthesis system to provide an audio database for Raramuri or Tarahumara, a Yuto-Nahua language spoken in Northern Mexico. In order to achieve the most natural speech possible, the synthesis system is proposed which uses a unit selection approach based on function words, suffix sequences (derivational and inflectional morphemes) and diphones of the language. In essence, the unknown suffix units were extracted from a corpus and recorded, along diphones and function words, in order to build the audio database that provides data for Text-to-Speech synthesis.

1 Introduction

The ultimate objective of Text-to-Speech (TTS) synthesis systems is to create applications which listeners, and users in general, cannot easily determine whether the speech he or she is hearing comes from a human or a synthesizer.

Synthesized speech can be produced by concatenating recorded units (waveforms) selected from a large, single-speaker speech database. The primary motivation for using a database with a large number of units that covers wider prosodic and spectral characteristics, gives us the great benefit to produce a synthesized speech that sounds more natural than those produced by systems that use a small set of controlled units (*e.g.* diphones) [1]. There is a paradigm for achieving high-quality synthesis that uses a large corpus of recorded speech units; it is called *unit-selection synthesis*. Unit selection is a method in which we can concatenate waveforms from different linguistic structures such as sentences, words, syllables, triphones, diphones and phones. Due to the increasing computer's storage capacity, we are able to create a corpus of prerecorded

Pronunciation Rules in Portuguese Regional Speech (PORT REG) for Coarticulation Process

Sara Candeias[1], Jorge Morais Barbosa [2]

[1] Instituto de Telecomunicações, Department of Computers and Electrical Engineering, University of Coimbra, PORTUGAL
[2] Departement of Portuguese Language, Faculty of Letters, University of Coimbra, PORTUGAL
saracandeias@co.it.pt, jbarbosa@ci.uc.pt

Abstract. This paper describes one aspect of an ongoing work to incorporate pronunciation variability in the Portuguese (PORT) speech system. This work focuses on the linguistic rules to improve the grapheme-(multi)phone transcription algorithm that will be implemented. Portuguese 'Beira Interior' regional speech (PORT-BI REG) is considered to be in the realm of coarticulation (post-lexical) phenomena. A set of linguistic rules for most of the common vowel transformation in an utterance (vocalic segments at both the left and right edges of the word) is presented. The analysis focuses on the distinctive features that originate vowel sound challenges in connected speech. The results are interesting from the point of view of setting up models to reconstruct a grapheme-phone transcription algorithm for Portuguese multi-pronunciation speech systems. We propose that the linguistic documentation of Portuguese minority speech can be an optimal start for Portuguese speech system development process, too.

Keywords: Text-to-Speech; coarticulation (phonology); structural analysis (linguistic features); pronunciation instruction (phonetic).

1 Introduction

Several frameworks have been proposed for the grapheme-to-phone transcription module for Portuguese language, such as [2, 3, 12]. However, the problem with the Portuguese regional speech under development is the shortage of speech and text corpora. This is one of the reasons why their linguistic structure has been very poorly investigated, especially at linguistic levels such as phonetics. The applications of the Portuguese speech system are mainly based on standard Portuguese language and on isolated word recognition. It is well known that the sequence of phones spoken by a human speaker is not the same sequence as that which derives from the phonetic transcription of a word in isolation. Coarticulation (post-lexical) rules must be included in the course of phonetic transcription. In order to obtain a more natural speech, these rules must be applied to varying sequences of phones. Several methods can be used to elicit grapheme-to-phoneme rules from pre-existing lexicons. However, these automatic techniques do not cope very well with the concurrent multi-

A Maximum Entropy (ME) Based Translation Model for Chinese Characters Conversion

Fai Wong[1], Sam Chao[1], Cheong Cheong Hao[1], and Ka Seng Leong[1]

[1] Faculty of Science and Technology of University of Macau,
Av. Padre Tomás Pereira S.J., Taipa, Macao
{derekfw, lidiasc}@umac.mo

Abstract. As the growth of exchange activities between four regions of cross strait, the problem to correctly convert between Traditional Chinese (TC) and Simplified Chinese (SC) is getting important and attention from many people, especially in business organizations and translation companies. Different from the approaches of many conventional code conversion systems, which rely on various levels of human constructed knowledge (from character set to semantic level) to facilitate the translation purpose, this paper proposes a Chinese conversion model based on Maximum Entropy (ME), a Machine Learning (ML) technique. This approach uses tagged corpus as the only information source for creating the conversion model. The constructed model is evaluated with selected ambiguous characters to investigate the recall rate as well as the conversion accuracy. The experiment results show that the proposed model is comparable to the state of the art conversion system.

Keywords: Maximum Entropy, Machine Learning, Natural Language Processing, Chinese translation, Traditional Chinese, Simplified Chinese.

1 Introduction

Modern Chinese typically involves two main dialects of writing, Traditional Chinese (TC) and Simplified Chinese (SC). In Chinese computing, these two systems adapt different coding schema for the computer to process the corresponding Chinese information. Traditional Chinese uses Big5 encoding while Simplified Chinese uses GB. For a Simplified Chinese document to be opened and read in a computer with Traditional Chinese operating system, conversion from Simplified Chinese encoding system into Traditional Chinese encoding is necessary for the purpose that the document can be further processed under the Traditional Chinese computer environment, and vice versa. As addressed by Wang [1] in the meeting of the 4th Chinese Digitization Forum, although there are many conversion systems implemented and available in the market, neither one of them can produce the conversion result with satisfaction. Reviewing the nature of this problem, Simplified Chinese is actually a simpler version of Traditional Chinese. It differs in two ways from the Traditional writing system: 1) a reduction of the number of strokes per character and 2) the reduction of the number of characters in use that is two different

Tracking Out-of-date Newspaper Articles

Frederik Cailliau[1,2], Aude Giraudel[3], Béatrice Arnulphy[4]

[1] Sinequa Labs, 12 Rue d'Athènes, F-75009 Paris
[2] LIPN-Univ. de Paris-Nord, 99 av. Jean-Baptiste Clément, F-93430 Villetaneuse
[3] Syllabs, 15 rue Jean-Baptiste Berlier, F-75013 Paris
[4] LIMSI-CNRS, BP 133, F-91403 Orsay*
cailliau@sinequa.com giraudel@syllabs.com beatrice.arnulphy@limsi.fr

Abstract. Local newspapers rely on their local correspondents to bring you the freshest news of your home town. Some of the articles written by these correspondents are not published immediately, but are put aside to be placed in a later edition. One of the challenges the editor in chief is confronted with, is to publish only up-to-date information about a local event. In this paper we present a system that tracks out-of-date newspaper articles to prevent their publishing. It firstly dates the event the article is talking about. The date detection grammar is written in a single but complex finite state automaton based on linguistic pattern matching. Secondly, it computes an absolute date for each relative date. A freshness score can then be deduced from the time difference between the extracted and the publication date. The system has been tested on several French local newspaper corpora. Baseline for the temporal extraction is our standard date extraction that was developed for general purposes.

Keywords: extraction of temporal information, named entities, events, information retrieval, newspaper articles

1 Introduction

Local daily newspapers rely on their network of correspondents to cover local events. Once revised by a journalist, these texts are able to be published as newspaper articles. Everyday the editor in chief validates or assembles the pages that will be published in the next edition. One of the challenges he is confronted with, is to validate only those articles covering hot news or coming events and to replace out-of-date articles by more recent ones. Unfortunately there is no easy or quick way to perform this task. To judge whether the article talks about a recent or a coming event, the editor has to read most of the article, since the day of writing is not a reliable indicator, when given. Even if most of the articles may be short, this activity is too time-expensive to be executed in the short delay before going to press. The editor's

* Some of the work in this article was done in collaboration with Béatrice Arnulphy during her internship at Sinequa in the summer of 2006 as a graduate student of the French university *Université de Provence*.

PtiClic: a game for vocabulary assessment combining JeuxDeMots and LSA

Mathieu Lafourcade[1], Virginie Zampa[2]

[1] LIRMM-TAL, Univ. Montpellier 2 France
mathieu.lafourcade@lirmm.fr
[2] LIDILEM-DIP Univ. Grenoble 3 France
virginie.zampa@u-grenoble3.fr

Abstract. One interesting path for designing software for vocabulary acquisition and assessment could be games. In this article we present PtiClic, a lexical game based on the principles behind JeuxDeMots and combined with LSA. Such a game can foster the interest of young people in developing lexical skills in either a tutored or an open environment.

Keywords: Lexical acquisition, lexical network, serious games, LSA, JeuxDeMots

1. Introduction

Developing software for vocabulary acquisition and/or assessment in general, and for young people in particular, is a risky business. There are several difficulties. First, we have to be able to design an activity that can foster an interest in learning which is not easy in the case of children. Then, the underlying dictionaries or lexical databases can prove tremendously hard to develop, especially if we want to go beyond just a list of words with parts-of-speech and venture into the realm of lexical functions and the relations intertwining terms.

Automated acquisition of lexical or functional relations between terms is necessary in a large number of tasks in Natural Language Processing (NLP) outscoping largely Technology-Enhanced Learning of language. These relations that we find generally in thesauruses or ontologies can of course be revealed in a manual way; for example, one of the oldest thesauruses is Roget's, its current version being (Kipfer, 2001), or the most famous lexical network is Wordnet (Miller, 1990). Such relations can be also determined computationally from corpora of texts, for example (Robertson and Spark Jones, 1976), (Lapata and Keller, 2005), or (Landauer et al 1998) in which statistical studies on the distributions of words are made. Moreover, many applications of NLP

Cascaded Regression Analysis based Temporal Multi-document Summarization

Ruifang He, Bing Qin, Ting Liu, Sheng Li
Information Retrieval Lab, Harbin Institute of Technology P.O.Box 321, HIT,P.R.China 150001
rfhe@ir.hit.edu.cn, http://ir.hit.edu.cn/

Keywords: temporal multi-document summarization, temporal semantic labeling, macro importance discriminative model, micro importance discriminative model

Received: February 5, 2009

Temporal multi-document summarization (TMDS) aims to capture evolving information of a single topic over time and produce a summary delivering the main information content. This paper presents a cascaded regression analysis based macro-micro importance discriminative model for the content selection of TMDS, which mines the temporal characteristics at different levels of topical detail in order to provide the cue for extracting the important content. Temporally evolving data can be treated as dynamic objects that have changing content over time. Firstly, we extract important time points with macro importance discriminative model, then extract important sentences in these time points with micro importance discriminative model. Macro and micro importance discriminative models are combined to form a cascaded regression analysis approach. The summary is made up of the important sentences evolving over time. Experiments on five Chinese datasets demonstrate the encouraging performance of the proposed approach, but the problem is far from solved.

Povzetek:

1 Introduction

Multi-document summarization is a technology of information compression, which is largely an outgrowth of the late twentieth-century ability to gather large collections of unstructured information on-line. The explosion of the World Wide Web has brought a vast amount of information, and thus created a demand for new ways of managing changing information. Multi-document summarization is the process of automatically producing a summary delivering the main information content from a set of documents about an explicit or implicit topic, which helps to acquire information efficiently. It has drawn much attention in recent years and is valuable in many applications, such as intelligence gathering, hand-held devices and aids for the handicapped.

Temporal multi-document summarization (TMDS) is the natural extension of multi-document summarization, which captures evolving information of a single topic over time. The greatest difference from traditional multi-document summarization is that it deals with the dynamic collection about a topic changing over time. It is assumed that a user has access to a stream of news stories that are on the same topic, but that the stream flows rapidly enough that no one has the time to look at every story. In this situation, a person would prefer to dive into the details that include the most important, evolving concepts within the topic and have a trend analysis.

The key problem of summarization is how to identify important content and remove redundant content. The common problem for summarization is that the information in different documents inevitably overlaps with each other, and therefore effective summarization methods are needed to contrast their similarities and differences. However, the above application scenarios, where the objects to be summarized face to some special topics and evolve with time, raise new challenges

Challenging Issues of Automatic Summarization: Relevance Detection and Quality-based Evaluation

Elena Lloret and Manuel Palomar
Department of Software and Computing Systems, University of Alicante
E-mail:{elloret, mpalomar}@dlsi.ua.es

Keywords: Natural Language Processing, Automatic Summarization, relevance detection, quality-based evaluation

Received:

This paper is about the Automatic Summarization task within two different points of view, focusing on two main goals. On the one hand, a study of the suitability for "The Code Quantity Principle" *in the Text Summarization task is described. This linguistic principle is implemented to select those sentences from a text, which carry the most important information. Moreover, this method has been run over the DUC 2002 data, obtaining encouraging results in the automatic evaluation with the ROUGE tool. On the other hand, the second topic discussed in this paper deals with the evaluation of summaries, suggesting new challenges for this task. The main methods to perform the evaluation of summaries automatically have been described, as well as the current problems existing with regard to this difficult task. With the aim of solving some of these problems, a novel type of evaluation is outlined to be developed in the future, taking into account a number of quality criteria in order to evaluate the summary in a qualitative way.*

Povzetek:

1 Introduction

The high amount of electronic information available on the Internet increases the difficulty of dealing with it in recent years. Automatic Summarization (AS) task helps users condense all this information and present it in a brief way, in order to make it easier to process the vast amount of documents related to the same topic that exist these days. Moreover, AS can be very useful for neighbouring Natural Language Processing (NLP) tasks, such as Information Retrieval, Question Answering or Text Comprehension, because these tasks can take advantadge of the summaries to save time and resources [1].

A summary can be defined as a reductive transformation of source text through content condensation by selection and/or generalisation of what is important in the source [2]. According to [3], this process involves three stages: *topic identification*, *interpretation* and *summary generation*. To identify the topic in a document what systems usually do is to assign a score to each unit of input (word, sentence, passage) by means of statistical or machine learning methods. The stage of interpretation is what distinguishes extract-type summarization systems from abstract-type systems. During interpretation, the topics identified as important are fused, represented in new terms, and expressed using a new formulation, using concepts or words not found in the original text. Finally, when the summary content has been created through abstracting and/or information extraction, it requires techniques of Natural Language Generation to build the summary sentences. When an extractive approach is taken, there is no generation stage involved.

Another essential part of the Text Summarization (TS) task is how to perform the evaluation of a summary. Methods for evaluating TS can be classified into two categories [4]. The first, intrinsic evaluations, test the summary on itself. The second, extrinsic evaluations, test how the summary is good enough to accomplish some other

Using Bagging and Boosting Techniques for Improving Coreference Resolution

Smita Vemulapalli
Center for Signal and Image Processing (CSIP),
School of Electrical and Computer Engineering, Georgia Institute of Technology,
Atlanta, GA 30332, USA
smita@ece.gatech.edu

Xiaoqiang Luo, John F. Pitrelli and Imed Zitouni
IBM T.J. Watson Research Center,
Yorktown Heights, NY 10598, USA
{xiaoluo,pitrelli,izitouni}@us.ibm.com

Keywords: Coreference resolution, information extraction, classifier combination, bagging, boosting, entity detection and tracking, majority voting

Received: February 5, 2009

Classifier combination techniques have been applied to a number of natural language processing problems. This paper explores the use of bagging and boosting as combination approaches for coreference resolution. To the best of our knowledge, this is the first effort that examines and evaluates the applicability of such techniques to coreference resolution. In particular, we (1) outline a scheme for adapting traditional bagging and boosting techniques to address issues, like entity alignment, that are specific to coreference resolution, (2) provide experimental evidence which indicates that the accuracy of the coreference engine can potentially be increased by use of multiple classifiers, without any additional features or training data, and (3) implement and evaluate combination techniques at the mention, entity and document level.

Povzetek:

1 Introduction

Classifier combination techniques have been applied to many problems in natural language processing (NLP). Popular examples include the ROVER system [Fiscus1997] for speech recognition, the Multi-Engine Machine Translation (MEMT) system [Jayaraman and Lavie2005], and also part-of-speech tagging [Brill and Wu1998, Halteren et al.2001]. Even outside the domain of NLP, there have been numerous interesting applications for classifier combination techniques in the areas of biometrics [Tulyakov and Govindaraju2006], handwriting recognition [Xu et al.1992] and data mining [Aslandogan and Mahajani2004] to name a few. Most of these techniques have shown a considerable improvement over the performance of single-classifier baseline systems and, therefore, lead us to consider implementing such a multiple classifier system for coreference resolution as well. To the best of our knowledge, this is the first effort that utilizes classifier combination techniques for improving coreference resolution.

This study shows the potential for increasing the accuracy of the coreference resolution engine by combining multiple classifier outputs and describes the combination techniques that we have implemented to establish and tap into this potential. Unlike other domains where classifier combination has been implemented, the coreference resolution application presents a unique set of challenges that prevent us from directly using traditional combination schemes [Tulyakov et al.2008]. We, therefore, adapt some of these popular yet

The Grammar of ReALIS and the Implementation of its Dynamic Interpretation

Gábor Alberti, Judit Kleiber
Department of Linguistics, University of Pécs, 7624 Pécs, Ifjúság 6., Hungary
realis@btk.pte.hu

Keywords: lexical and discourse semantics, Hungarian, total lexicalism

Received: [Enter date]

ReALIS, REciprocal And Lifelong Interpretation System, is a new "post-Montagovian" theory concerning the formal interpretation of sentences constituting coherent discourses, with a lifelong model of lexical, interpersonal and encyclopedic knowledge of interpreters in its center including their reciprocal knowledge on each other. Section 2 provides a 2 page long summary of its 40 page long mathematical definition [4]. Then we show the process of dynamic interpretation of a Hungarian sentence (Hungarian is a "challenge" because of its rich morphology, free word order and sophisticated information structure). We show how an interpreter can anchor to each other in the course of dynamic interpretation the different types of referents occurring in copies of lexical items retrieved by the interpreter on the basis (of the morphemes, word order, case and agreement markers) of the sentence performed by the speaker. In Section 4 the computational implementation of ReALIS is demonstrated.

1 Introduction

ℜeALIS [2] [4], *REciprocal And Lifelong Interpretation System*, is a new "post-Montagovian" [15] [17] theory concerning the formal interpretation of sentences constituting coherent discourses [9], with a *lifelong* model [1] of lexical, interpersonal and cultural/encyclopedic knowledge of interpreters in its center including their *reciprocal* knowledge on each other. The decisive theoretical feature of ℜeALIS lies in a peculiar reconciliation of three objectives which are all worth accomplishing in formal semantics but could not be reconciled so far.

The first aim concerns the exact formal basis itself ("Montague's Thesis" [20]): human languages can be described as interpreted *formal* systems. The second aim concerns *compositionality*: the meaning of a whole is a function of the meaning of its parts, practically postulating the existentence of a *homomorphism* from syntax to semantics, i.e. a rule-to-rule correspondence between the two sides of grammar.

In Montague's interpretation systems a traditional logical representation played the role of an intermediate level between the syntactic representation and the world model, but Montague argued that this intermediate level of representation can, and should, be eliminated. (If α is a compositional mapping from syntax to discourse representation and β is a compositional mapping from discourse to the representation of the world model, then $\gamma=\alpha^{\circ}\beta$ must be a compositional mapping directly from syntax to model.) The post-Montagovian history of formal semantics [17] [9], however, seems to have proven the opposite, some principle of "discourse representationalism": "some level of [intermediate] representation is indispensable in modeling the interpretation of natural language" [14].

The Thesis of ℜeALIS is that the two fundamental Montagovian objectives *can* be reconciled with the principle of "discourse representationalism" – by embedding discourse representations in the world model, getting rid of an intermediate level of representation in this way while preserving its content and relevant structural characteristics. This idea can be carried out in the larger-scale framework of embeddig discourse representations in the world model *not directly* but as parts of the representations of interpreters' minds, i.e. that of their (permanently changing) information states [3].

2 Definition

The frame of the mathematical definition of ℜeALIS (whose 40 page long complete version is available in [4] (Sections 3-4)) is summarized in this section. As interpreters' mind representations are part of the WORLD MODEL, the definition of this model $\mathfrak{R} = \langle U, W_0, W\rangle$ is a quite complex structure where

- U is a countably infinite set: the UNIVERSE
- $W_0 = \langle U_0, T, S, I, D, \Omega, A\rangle$: the EXTERNAL WORLD
- W is a partial function from set I×Tm where W[i,t] is a quintuple $\langle U[i], \sigma[i,t]^{\Pi}, \alpha[i,t]^{\Psi}, \lambda[i,t]^{\Lambda}, \kappa[i,t]^{K}\rangle$: the INTERNAL-WORLD FUNCTION.

The external world consists of the following components:

- U_0 is the external universe ($U_0 \subset U$), whose elements are called entities
- $T = \langle T, \Theta\rangle$ is a structured set of temporal intervals ($T \subset U_0$)
- $S = \langle S, \Xi\rangle$ is a structured set of spatial entities ($S \subset U_0$)
- $I = \langle I, Y\rangle$ is a structured set of interpreters ($I \subset U_0$)

Classifying Library Titles Based Solely on Their Title*

Ricardo A. Argüelles[1], Hiram Calvo[1,2], Salvador Godoy[1], Alexander Gelbukh[1]

[1]Center for Computing Research, National Polytechnic Institute
Mexico City, 07738, Mexico
ravila06@sagitario.cic.ipn.mx; hcalvo@cic.ipn.mx;
sgodoyc@cic.ipn.mx, gelbukh@gelbukh.com

[2]Nara Institute of Science and Technology, Takayama, Ikoma, Nara 630-0192, Japan
calvo@is.naist.jp

Keywords: Library classification, LCC, Scarce information classification, logical-combinatorial methods

Received: 2009.02.04

Abstract. *Many books comply with the Library of Congress Classification (LCC) by adding their classification number in the first pages. This is useful for many libraries worldwide because it makes it possible to search and retrieve books by type of content and it has now become a standard. However, not every book has been pre-classified; particularly, in many universities, new dissertations have to be classified manually. Although there are many systems available for automatic classification, all of them use additional information to the title of the book. In this work, we experiment with the classification of new books by using only their title, which would allow indexing books massively in an automatic fashion. We propose a new measure for comparison, which mixes two well known classification techniques: A Lesk voting scheme and Term Frequency in documents (TF). In addition, we experiment with different weighing as well as Logical-Combinatorial methods such as ALVOT in order to determine the contribution of the title in its correct classification. We found this contribution to be around one third, as we correctly classified 36% (average from each branch) of 122,431 previously unseen titles (in total) trained with 489,726 samples (in total) of one major branch (Q) of the LCC catalog.*

1 Introduction

The Library of Congress Classification (LCC) is widely known and used by many important libraries internationally [6]. It is the system with the widest coverage of books. One of the most important tasks of librarians is book classification. A classification system designed to meet their requirements is the LCC. Besides using the previously-assigned LCC for each book, librarians need to classify other works such as dissertations, articles, magazines, which in most cases lack a previously-assigned LCC [7]. We focus our work on creating an algorithm to automatically assign a classification based only on the most basic piece of information available—the title of the publication. We explore the level of attainment that it is possible to obtain given this strong restriction. We faced several problems, such as similar titles in different classes, a noisy data set. We tested using 5 algorithms, some of them are very simple, and others, such as those based on Logical Combinatorial methods are more complex. In the following section we elaborate on similar works. In Section 3 we explain the different algorithms presented. In Section 4 we explain our experiments with the LCC catalog and results; and finally, in Section 5 we draw our conclusions and outline our future work.

2 Related Work

Table 1 summarizes previous work for book classification and compares the information they use with regard to the information that we use.

1. Predicting Library of Congress Classification from Library of Congress Subject Headings, Frank & Paynter, 2004 [3]
2. The Utility of Information Extraction in the classification of books., Betts *et. al*., 2007 [4]
3. Experiments in Automatic Library of congress Classification, Larson, 1992 [5]
4. Challenges in automated classification using library classification schemes—Pharos, Kwan Yi, 2006 [2]

The last column represents the characteristics of our work.

Next, we explain LCSH and MARC.

1. LCSH (*Library of congress subject headings*) is a collection of synonyms and antonyms of several terms related with book contents. This collection is updated by the Library of Congress. LCSH is widely used for book searching where searches such

* We thank the support of the Mexican Government (SNI, SIP-IPN, COFAA-IPN, and PIFI-IPN) and the Japanese Government. The second author is a JSPS fellow.

Low-Bias Extraction of Domain-Specific Concepts

Axel-Cyrille Ngonga Ngomo
University of Leipzig, Johannisgasse 26, Leipzig D-04103, Germany,
ngonga@informatik.uni-leipzig.de,
WWW home page: `http://bis.uni-leipzig.de/AxelNgonga`

Keywords: Natural Language Processing, Local Graph Clustering, Knowledge-Free Concept Extraction

Received: February 5, 2009

The availability of domain-specific knowledge models in various forms has led to the development of several tools and applications specialized on complex domains such as bio-medecine, tourism and chemistry. Yet, most of the current approaches to the extraction of domain-specific knowledge from text are limited in their portability to other domains and languages. In this paper, we present and evaluate an approach to the low-bias extraction of domain-specific concepts. Our approach is based on graph clustering and makes no use of a-priori knowledge about the language or the domain to process. Therefore, it can be used on virtually any language. The evaluation is carried out on two data sets of different cleanness and size.

Povzetek:

1 Introduction

The recent availability of domain-specific knowledge models in various forms has led to the development of information systems specialized on complex domains such as bio-medecine, tourism and chemistry. Domain-specific information systems rely on domain knowledge in forms such as terminologies, taxonomies and ontologies to represent, analyze, structure and retrieve information. While this integrated knowledge boosts the accuracy of domain-specific information systems, modeling domain-specific knowledge manually remains a challenging task. Therefore, considerable effort is being invested in developing techniques for the extraction of domain-specific knowledge from various resources in a semi-automatic fashion. Domain-specific text corpora are widely used for this purpose. Yet, most of the current approaches to the extraction of domain-specific knowledge in the form of terminologies or ontologies are limited in their portability to other domains and languages. The limitations result from the knowledge-rich paradigm followed by these approaches, i.e., from them demanding hand-crafted domain-specific and language-specific knowledge as input. Due to these constraints, domain-specific information systems exist currently for a limited number of domains and languages for which domain-specific knowledge models are available. An approach to remedy the high human costs linked with the modeling of domain-specific knowledge is the use of low-bias, i.e., knowledge-poor and unsupervised approaches. They require little human effort but more computational power to achieve the same goals as their hand-crafted counterparts.

In this work, we propose the use of low-bias approaches for the extraction of domain-specific terminology and concepts from text. Especially, we study the low-bias extraction of concepts out of text using a combination of metrics for domain-specific multi-word units and graph clustering techniques. The input for this approach consists exclusively of a domain-specific text corpus. We use the Smoothed Relative Expectation [9] to extract domain-specific multi-word units from the input data set. Subsequently we use SIGNUM [10] to compute a domain-specific lexicon. Finally, we use BorderFlow, a novel general-purpose graph clustering algorithm, to cluster the domain-specific terminologies to concepts. Our approach is unsupervised and makes no use of

Automatic Identification of Lexical Units

Vidas Daudaravicius
Vytautas Magnus University, Faculty of Informatics
Vileikos 8, Kaunas, Lithuania
E-mail: vidas@donelaitis.vdu.lt

Keywords: lexical unit, lexical unit identification, token/type ratio, dice score, corpus size, average minimum law

Received:

Lexical unit is a word or collocation. Extracting lexical knowledge is an essential and difficult task in NLP. The methods of extracting of lexical units are discussed. We present a method for the identification of lexical boundaries. The problem of necessity of large corpora for training is discussed. The advantage of identification of lexical boundaries within a text over traditional window method or full parsing approach allows to reduce human judgment significantly.

Povzetek:

1 Introduction

Identification of a lexical unit is an important problem in many natural language processing tasks and refers to the process of extracting of meaningful word chains. The Lexical unit is a fuzzy term embracing a great variety of notions. The definition of the lexical unit differs according to the researchers interests and standpoint. It also depends on the methods of extraction that provide researchers with lists of lexical items. Most lexical units are usually single words or constructed as binary items consisting of a node and its collocates found within a previously selected span. The lexical unit can be: (1) a single word, (2) the habitual co–occurrence of two words and (3) also a frequent recurrent uninterrupted string of words. Second and third notion refers to the definition of a collocation or a multi–word unit. It is common to consider a single word as a lexical unit. A big variety of the definition of the collocation is presented in Violeta Seretan work [12]. Fragments of corpus or strings of words consisting of collocating words are called collocational chains [7]. For many years the final agreed definition of the collocation is not made. Many syntactical, statistical and hybrid methods have been proposed for collocation extraction [13], [1], [5], [4]. In [10], it is shown that MWEs are far more diverse and interesting than is standardly appreciated. MWEs constitute a key problem that must be resolved in order for linguistically precise NLP to succeed. Although traditionally seen as a language independent task, collocation extraction relies nowadays more and more on the linguistic preprocessing of texts prior to the application of statistical measures. In [14] it is provided a language-oriented review of the existing extraction work.

In our work we compare Dice and Gravity Counts methods for the identification of lexical units by applying them under the same conditions. The definition of what is a Lexical Unit in a linguistic sence is not discussed in this paper. New presented technique extracts collocations like 'in the' that do not have meaning and have functional purpose. A question of keeping such collocations as lexical units is left open. At the same time, it is interesting to see that the frequency lists of such lexical units for English and Lithuanian (memeber of Balto-Slavonic language group) are now comparable.

2 Extracting vs. Abstracting

Most of the collocation definitions refer to the collocation, which is constructed in an abstracting way. The collocations are not gathered directly from the text but rather constructed using syntactic and statistical information. The abstracted

Disentangling the Wikipedia Category Graph for Corpus Extraction

Axel-Cyrille Ngonga Ngomo, Frank Schumacher

***Abstract*—In several areas of research such as knowledge management and natural language processing, domain-specific corpora are required for tasks such as terminology extraction and ontology learning. The presented investigations herein are based on the assumption that Wikipedia can be used for the purpose of corpus extraction. It presents the advantage of possessing a semantic layer, which should ease the extraction of domain-specific corpora. Yet, as the Wikipedia category graph is scale-free, it can not be used as it is for these purposes. In this paper, we propose a novel approach to graph clustering called BorderFlow, which we use and evaluate on the Wikipedia category graph. Additional possible applications of these results in the area of information retrieval are presented.**

***Index Terms*—Natural Language Processing, Local Graph Clustering, Corpus Extraction**

I. Introduction

SEVERAL areas of research (e.g., knowledge management, natural language processing (NLP)) require domain-specific knowledge for tasks such as information retrieval (IR), lexicon extraction and ontology learning. In order to remedy the lack of domain-specific text corpora in certain domains, the Web has been used as supplementary data source. The investigations presented herein are based on the assumption that Wikipedia can provide a good starting point for this task, as it provides high-quality text and is freely accessible. A naive approach to the extraction of domain-specific text corpora from Wikipedia would consist of two steps: selecting the node(s) of the category graph which describe best the data required, and fetching iteratively all related categories (related means here, for example, categories appearing in the same articles or subcategories). Yet, such an approach would fail due to the fact that the Wikipedia category graph (WCG) presents a high degree of connectivity as it is scale-free [12]. An iterative approach would thus select too many if not all categories when iterated sufficiently often, since it would tend to integrate hubs. Furthermore, the WCG does not present an explicit similarity relation between categories, which could be used for the purpose described above. In this paper, we present a novel soft graph clustering approach, BorderFlow, which allows the discovery of clusters of paradigmatically related categories. Consequently, it enables the retrieval of domain-specific corpora, which can be extracted by retrieving all the pages tagged with the categories belonging to a certain domain, i.e., to a certain cluster. It is of great importance that the algorithm is fuzzy, as a category can belong to more than one domain. For example, "graph clustering" can be seen as belonging to mathematics and to computer sciences. BorderFlow allows the online detection (i.e. the detection at runtime) of clusters in large graphs generated by a given seed, making it suitable to be used in several Web2.0 applications such as the instantiation and exploration of taxonomies and ontologies and the generation of novel user interfaces for adaptive IR.

The rest of this paper is organized as follows: the next section presents some related work on graph clustering. In the subsequent section, the theoretical background of BorderFlow is elucidated, including a heuristic guaranteeing short run times on large graphs such as the WCG. Thereafter, the current implementation is described. The results achieved on the WCG and further possible applications of this clustering algorithm in the context of NLP are finally discussed.

Axel-Cyrille Ngonga Ngomo and Frank Schumacher are with the Department of Computer Science, University of Leipzig, Germany, Johannisalle 23, Room 5-22, 04103 Leipzig, e-mail: ngonga@informatik.uni-leipzig.de

Manuscript received February 5, 2009; revised XX

II. Related work

Graph clustering algorithms have been a topic of intense research in the past decade. They try to maximize or minimize a given criterion such as conductance, inter-cluster similarity or silhouette factor [9]. Markov Clustering (MCL) [11] for example tries to maximize the flow within a cluster. It is based on the idea that random walks that visits a cluster are likely not to leave the cluster until they have visited many of its vertices. Using a combination of inflation and expansion operators on all elements of the adjacency matrix, the algorithm generates a partition of the graph vertices.

Another clustering algorithm, which uses global information is the Iterative Conductance Cutting (ICC) algorithm [6]. The underlying idea of the algorithm is to iteratively separate clusters by finding minimal conductance cuts. The algorithm is NP-hard by itself, although it can be made polynomial when using a heuristic based on the eigenvalue of the adjacency matrix.

A popular algorithm in the area of NLP is Pantel's Clustering By Committee (CBC, [10]). It is a two-step algorithm, which first discovers unambiguous cluster centers (so-called committees) by computing sub-clusters in the top-k similarity graph generated out a complete similarity graph. Committees maximize the intra-cluster similarity while minimizing the inter-cluster similarity. The elements which do not belong to any committee are subsequently clustered in a fuzzy fashion in the second pass. CBC demands the setting of the parameter k, which can lead to too strict/loose committees and thus to an inadequate clustering of the graph at hand. Nevertheless CBC

Semantic Enterprise Search

(but no Web 2.0)

Ronald Winnemöller
University of Hamburg
ronald.winnemoeller@uni-hamburg.de

***Abstract*—In this paper, we propose semantic enterprise search as promising technical methodology for improving on accessibility to institutional knowledge. We briefly discuss the nature of knowledge and ignorance in respect to web-based information retrieval before introducing our particular view on semantic search as tight fusion of search engine and semantic web technologies, based on semantic annotations and the concept of intra-institutionwise distributed extensibility while still maintaining free keyword search functionality. Consequently , our architecture implementation makes strong use of the Aperture and Lucene software frameworks but introduces the novel concept of "RDF documents". Because our prototype system is not yet complete, we are not able to provide performance statistics but instead we present a concise example scenario.**

I. Introduction

Intuitively, it is the duty of Universities (and, to a certain degree, of technikons and other schools) to produce knowledge in research and teaching.

This, we might assume, is what they do very well.

Unfortunately, we may also find that keeping, consolidating and making accessible that knowledge, even when we restrict ourselves to electronically stored knowledge, is a field that is neglected in many cases – a fact that is acknowledged by several institutions as stated in the *Implementation of the Berlin Declaration on Open Access*, cf. [1] and the *Berlin Declaration on Open Access to Knowledge in the Sciences and Humanities* itself, cf. [2].

We can, for example, identity the following "knowledge leaks" as common for many institutions:

- The conventional, if not required publishing process often means that a researcher sums up his knowledge in an journal article or conference paper – which eventually gets published by some commercial company. It does not necessarily mean that this publication is kept (electronically) in the realm of the home university of the mentioned researcher and be available to other members – students or fellow researchers – of that institution. This issue is also reflected on in [1].
- Many institutions run web-accessible publication databases, like eprint archives, citation indexes and such – sometimes even at the department or team level. While these repositories usually provide for intra-repository search functionality, they cannot be searched using an institution-wise global methodology because of the well known "hidden web" problem (cf. e.g. [3], [4]).
- ELearning, content management and "Web 2.0" systems (such as departmental wikis, blogs, etc.) usually require authentication (and authorization) prior to accessing content. These knowledge sources cannot be accessed through global methods without special adaption to these requirements. Even then, automatic retrieval methods still face the above mentioned "hidden web" problem.
- Other types of knowledge material are transmitted electronically but are of transient nature, such as RSS-Feeds (on an organizational basis), mails, filestores, etc. We hesitate to call these types "documents" because of their temporary character - but nevertheless they might contain valuable knowledge nonetheless.
- Some publications are simply not available electronically, because they were published through a "pure-paper"[1] process.
- Certain documents may be accessible through the "visible" intraweb of an institution, but due to an ineffective implementation of the retrieval process they may still be inaccessible.

In this paper, we propose a *technical* methodology for improving on accessibility to institutional knowledge.

We will not, however, try to solve *social* institutional issues such as implementing open access publishing processes, etc[2].

The remainder of this paper is structured as follows:

In the next section, we will briefly discuss the term "knowledge". After this, we will put our work in an appropriate scientific context in section III, followed by a description of our own approach (section IV), including our proposed architecture and already implemented modules. Since we are still working on a concise evaluation procedure, we will instead preliminarily provide a realistic scenario in section IV-C as indication for the intended functionality. Subsequently, we conclude in section V

II. A Brief Elaboration on a search-related Notion of "Knowledge"

In order to clarify what we are talking about, we need to discuss what we mean when using the term "knowledge". Unfortunately, even a superficial definition of the nature of "knowledge" is far beyond the scope of a paper like this one but we would like to prevent some common misunderstandings:

[1] As opposed to "paper-less".

[2] We will rather leave this issue to the Web2.0 community ...

Semantic Web Framework for Very Large Ontologies Development

Sergey Yablonsky[1]

[1]Graduate School of Management, St. Petersburg State University,
Information Technologies in Management Dpt.,
Volkhovsky Per. 3, St.-Petersburg, 199004, Russia,
Email: serge_yablonsky@hotmail.com

Abstract. This paper deals with the development of the Semantic Web framework for very large ontologies. The Semantic Web is often associated with specific XML-based standards for semantics, such as RDF and OWL. Application of the lexical ontologies such as WordNet and others for different tasks on the Semantic Web requires a representation of them in RDF and/or OWL with possibility of the different ontology mappings, semantic workflows, services and other semantic technologies.

Keywords: Semantic Web, OWL, RDF, Resource Description Framework

1 Introduction

The Semantic Web, a Web with the meaning**,** is often associated with specific XML-based standards for semantics, such as RDF and OWL [http://www.w3.org/RDF/, http://www.w3.org/TR/owl-features/]. If HTML and the Web made all the online documents look like one huge book, RDF, schema, and inference languages will make all the data in the world look like one huge database [1]. The Semantic Web Layer Cake (Fig.1) shows that there are different layers in the Semantic Web and that they do different things. Some of the layers can take different forms. Each of the layers is less general than the layers below.

RDF (Resource Description Framework) is a markup language for describing information and resources on the web. RDF represents data as a set of statements consisting of a 'subject', a 'predicate', and an 'object'. Each statement is also known as a 'triple' or a 'relationship'. The Subject and the Predicate are named resources. A resource is represented by a URI. The Object can be a literal or another resource.

Example of RDF data:

```
   (Subject)      (Predicate)  (Object)
  <SergeyYablonsky><name>       "Serge Yablonsky".
  <SergeyYablonsky><email>      "serge_yablonsky@hotmail.com".
  <SergeyYablonsky><PhDAdviser><AndreySukhonogov>.
 <AndreySukhonogov><email>      "ASukhonogov@rambler.ru>.
```

Putting information into RDF files, makes it possible for computer programs ("web spiders") to search, discover, pick up, collect, analyze and process information from the web. The Semantic Web uses RDF to describe web resources.

Nowadays there exists a linked set of different Semantic Web resources as it is shown on the Fig.2. On Fig.3 the Linking Open Data (LOD) Constellation is shown.

SMM: Detailed, Structured Morphological Analysis for Spanish

Cerstin Mahlow and Michael Piotrowski

***Abstract*—We present a morphological analyzer for Spanish called *SMM*. SMM is implemented in the grammar development framework *Malaga*, which is based on the formalism of *Left-Associative Grammar*. We briefly present the Malaga framework, describe the implementation decisions for some interesting morphological phenomena of Spanish, and report on the evaluation results from the analysis of corpora. SMM was originally only designed for analyzing word forms; in this article we outline two approaches for using SMM and the facilities provided by Malaga to also generate verbal paradigms. SMM can also be embedded into applications by making use of the Malaga programming interface; we briefly discuss some application scenarios.**

***Index Terms*—Natural language processing, morphology, Malaga, Spanish**

I. Introduction

MORPHOLOGY is one of the core processes of language. By applying the rules for inflection, derivation, and compounding, humans are able to create and understand the word forms required to communicate, including the creation of new words from existing words. To understand an utterance in some language we have to know the rules of syntax and morphology, as these are essential prerequisites for dealing with semantics or even pragmatics.

From the point of view of computational linguistics, morphological resources form the basis for all higher-level applications. A morphological component should thus be capable of analyzing single word forms as well as whole corpora, and it should provide detailed analyses describing the relevant morphological processes. For evaluation purposes, it should also provide statistical information on speed, accuracy, etc. when analyzing large corpora.

The *Malaga* system provides a framework that supports both the development of morphological components and their application. In section II, we will give a short overview of the Malaga framework and the underlying formalism of *Left-Associative Grammar*. In the rest of this article, we will then present a specific application of Malaga, a morphological component for Spanish – the *Spanish Malaga Morphology (SMM)*.

In section III, we describe some important morphological phenomena of Spanish and present a number of principles for handling these phenomena, which guided the design of SMM. In section IV, we describe the implementation of SMM. Section V reports on the performance of SMM on two corpora. This is followed by an overview of related work (section VI) and a discussion of the use of SMM in a variety of applications. Section VIII summarizes the properties and specific advantages of SMM and outlines future work.

The authors are with the Institute of Computational Linguistics, University of Zurich, Binzmühlestrasse 14, 8050 Zurich, Switzerland (e-mail: mahlow@cl.uzh.ch; mxp@cl.uzh.ch).

II. Malaga and Left-Associative Grammar

Malaga is a software package for the development and application of morphology and syntax grammars based on the *Left-Associative Grammar* (LAG) formalism [1], providing a specialized programming language and associated development tools.

Left-Associative Grammar is based on non-deterministic finite automata. As implemented in Malaga, the analysis states are augmented by arbitrarily complex feature structures. In a morphology grammar, the symbols read from the input are allomorphs. The feature structures allow to store all available information about the involved allomorphs and the values resulting from the concatenation of these allomorphs. For the presentation of analysis results the information can be filtered to show only the features needed for a certain purpose.

Morphological components implemented in Malaga are based on the *allomorph approach*, which we will briefly describe in section IV-A.[1] Thus, the run-time lexicon used by Malaga grammars is an *allomorph lexicon* generated from a base form lexicon by applying allomorphy rules at compile time.[2]

Malaga is able to process text in UTF-8 encoding. Besides the morphological component for Spanish described in this paper, a number of Malaga grammars for morphological and syntactical analysis of English, Finnish, German, Italian, and Korean have been created, both at the University of Erlangen (Germany), where Malaga was originally developed, and elsewhere.

Malaga is freely available under the GNU Public License (GPL). For the work described in this paper we used Malaga version 7.12 on Mac OS X and Linux.[3]

III. Spanish Morphology

Spanish, an Ibero-Romance language, is one of the most widely-spoken languages of the world. On the grounds of its rich verbal morphology it can be classified as an inflecting language; however, almost all of the noun inflections have disappeared, with only a plural marker remaining.

In this section, we will give a short overview of morphological processes and phenomena of Spanish, and briefly describe orthographical issues. We will present them in a way that allows us to define principles for the implementation of SMM.

[1] See [2] for a comparison of methods for morphological analyzers.

[2] See Björn Beutel: *Malaga. A Grammar Development Environment for Natural Languages*, http://home.arcor.de/bjoern-beutel/malaga/ [last access 2009-02-04].

[3] We have also used this and earlier versions of Malaga on various versions of Solaris, HP-UX, and NetBSD.