Has Computational Linguistics Become More Applied?

Kenneth Church

One Microsoft Way
Redmond WA 98052 USA
church@microsoft.com

Abstract. Where the field has been and where it is going? It is relatively easy to know where we have been, but harder (and more valuable) to know where we are going. The title of this paper, borrowed from Hull, Jurafsky and Martin (2008), suggests that applications have become more important, and that industrial laboratories will become increasingly prestigious.

1 Rise of Statistical Methods

Some trends are pretty well established. The Association for Computational Linguistics (ACL) is clearly accepting more statistical papers than it used to. Both Bob Moore (personal communication) and Fred Jelinek (personal communication) performed independent surveys and found a dramatic increase in statistical papers over the last couple decades [7].

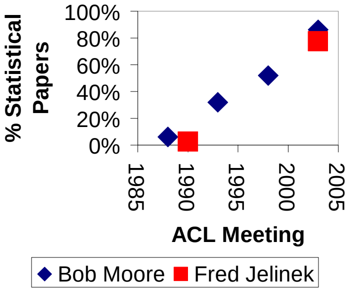


Figure 1. Rise of Statistical Methods

Hull, Jurafsky and Martin [9], henceforth HJM, came to a similar conclusion by applying sophisticated modern methods such as LDA (Latent Dirichlet Allocation [4]) to a corpus of 14,000 documents from the ACL Anthology [3]. HJM suggest that 1988 was a particularly important year. I will refer to their suggestion as the "big bang." HJM called out two papers from 1988. It is probably not an accident that both of these papers came from well-funded industrial laboratories.

1. IBM: Brown *et al*, 1988 [5], a seminal paper on Statistical Machine Translation
2. AT&T Bell Labs: Church, 1988 [6], Part of Speech Tagging

Opportunities for Natural Language Processing Research in Education

Jill Burstein

Educational Testing Service
Rosedale Road, MS 12R
Princeton, New Jersey 08540
USA, jburstein@ets.org

Abstract. This paper discusses emerging opportunities for natural language processing (NLP) researchers in the development of educational applications for writing, reading and content knowledge acquisition. A brief historical perspective is provided, and existing and emerging technologies are described in the context of research related to content, syntax, and discourse analyses. Two systems, e-rater® and *Text Adaptor*, are discussed as illustrations of NLP-driven technology. The development of each system is described, as well as how continued development provides significant opportunities for NLP research.

Keywords: natural language processing, automated essay scoring and evaluation, text adaptation, English language learning, educational technology

1 Introduction

Theoretically, opportunities for natural language processing (NLP) research have existed in education in the area of reading research since the 1940's; writing research since the 1960's, and in the teaching of content knowledge since the early 1970's. While opportunities have existed for several decades, the general lack of computer-based technology proved to be an obstacle early on. In later years, when computers became increasingly more available, the lack of well-instantiated technological infrastructure (especially in schools), where educational applications could be broadly distributed and used, presented yet another obstacle – however, we can still see where the opportunities existed and how they grew over time.

Research in the area of *readability*, or *text quality* investigates the linguistic aspects of text that make a text relatively easier or more difficult to comprehend or

Information Structure in a Formal Framework: Unification-based Combinatory Categorial Grammar

Maarika Traat

University of Tartu, Estonia
Institute of Computer Science
maarika.traat@ut.ee

Abstract. This paper presents Unification-based Combinatory Categorial Grammar (UCCG): a grammar formalism that combines insights from Combinatory Categorial Grammar with feature structure unification. Various aspects of information structure are incorporated in the compositional semantics. Information structure in the semantic representation is worked out in enough detail to allow for the determination of accurate placement of pitch accents, making the representation a suitable starting point for speech generation with context appropriate intonation. UCCG can be used for parsing and generating prosodically annotated text, and uses a semantic representation that is compatible with the currently available 'off-the-shelf' automatic inference tools. As such the framework has the potential to advance spoken dialogue systems.

1 Introduction

Information structure plays a crucial role in ensuring the coherence of a text or discourse. As such, its incorporation could improve the performance of a variety of natural language applications. Yet actual computational systems ignore it almost entirely. There are two main obstacles for its inclusion. Research into information structure tends to concentrate on specific linguistic phenomena, while its overall effect on compositional semantics applicable for a range of sentences is rarely worked out in enough detail to be useful for computational implementation. The second problem is that the formalizations that describe the semantic impact of information structure tend to use higher-order logic [1–3], which limits the use of inference in practice [4]. There is a tight connection between information and intonation. Thus, the inclusion of information structure in a formal grammar framework paves the way for improvements in the quality of intonation in systems where output is generated from semantic representations.

This paper presents Unification-based Combinatory Categorial Grammar (UCCG), which integrates aspects of Combinatory Categorial Grammar [3], Unification Categorial Grammar [5, 6], and Discourse Representation Theory [7]. It offers a compositional analysis of information structure and a semantics compatible with first-order logic. The use of feature structure unification to combine grammatical categories makes UCCG easy to implement computationally, and allows for the integration of prosodic information in the semantics in a transparent and systematic way. UCCG provides a link between prosodically annotated

A *karaka* based annotation scheme for English

Ashwini Vaidya, Samar Husain, Prashanth Mannem, Dipti Misra Sharma

Language Technologies Research Centre, International Institute of Information Technology, Hyderabad, India
{ashwini_vaidya, samar, prashanth}@research.iiit.ac.in, dipti@iiit.ac.in

Abstract. The paper describes an annotation scheme for English based on Panini’s concept of karakas. We describe how the scheme handles certain constructions in English. By extending the karaka scheme for a fixed word order language, we hope to bring out its advantages as a concept that incorporates some ‘local semantics’. Our comparison with PTB-II and PropBank brings out its intermediary status between a morpho-syntactic and semantic level. Further work can show how this could benefit tasks like semantic role labeling and automatic conversion of existing English treebanks into this scheme.

1 Introduction

Beginning with the Penn treebank [14], treebank annotation has remained an important research area in CL and NLP. The PTB itself has become richer by incorporating various facets of language phenomenon over the basic phrase structure syntactic representation. Some of these include addition of grammatical relations (PTB-II, [15], [14]), predicate argument structure (PropBank [11]), and immediate discourse structure (PDTB [16]). Treebanks in other languages have continued to enrich this research initiative. For morphologically rich languages like Czech, one major effort has been the Prague Dependency Treebank [8], which has used a dependency based formalism. The Hyderabad dependency treebank- HyDT [1] for Hindi also follows the dependency based approach. In this paper we elaborate & extend the karaka based annotation scheme used in HyDT to English. We also compare some of our tags with similar tags in other well known schemes. As we will see from the examples discussed in the paper, karaka relations capture some level of ‘local semantics’. As Rambow et al. [19] state, “local semantic labels are relevant to the verb meaning in question, while global semantic labels are relevant across different verbs and verb meanings”. Previous work [18] has used an annotation scheme based on a dependency structure for English but our scheme differs considerably.

The paper is arranged as follows; in Section 2 we discuss the concept of karaka relations. Section 3 describes the data used for annotation. In Section 4 we explain the tagset used. We show how some English constructions are handled in the scheme in Section 5. Section 6 compares our work with a dependency version of Penn Treebank as well as with PropBank. We discuss some related issues in Section 7.

Substring Statistics

Kyoji Umemura[1] and Kenneth Church[2]

[1] Toyohashi University of Technology, Tempaku, Toyohashi, Aichi 441-8580, Japan
[2] Microsoft, One Microsoft Way, Redmond, WA 98052, USA

Abstract. The goal of this work is to make it practical to compute corpus-based statistics for all substrings (ngrams). Anything you can do with words, we ought to be able to do with substrings. This paper will show how to compute many statistics of interest for all substrings (ngrams) in a large corpus. The method not only computes standard corpus frequency, $freq$, and document frequency, df, but generalizes naturally to compute, $df_k(str)$, the number of documents that mention the substring str at least k times. df_k can be used to estimate the probability distribution of str across documents, as well as summary statistics of this distribution, e.g., mean, variance (and other moments), entropy and adaptation.

1 Introduction

Substring (ngram) statistics are fundamental to nearly everything we do: language modeling (for speech recognition, OCR and spelling correction), compression, information retrieval, word breaking and more. Many textbooks discuss applications of ngrams including [14] [16] [17] [13] [8] [5] [7] [12]. This paper describes an easy-to-follow procedure for computing many popular statistics for all substrings (ngrams) in a large corpus. **C** code is posted at [21].

[19] showed how to compute standard corpus frequency, $freq$, and document frequency, df, for all substrings in a large corpus. Document frequency is a commonly used statistic, especially in the Information Retrieval community [2] [5]. Document frequency is traditionally defined over words or terms, though we will apply it to substrings.

Definition 1. $df(str) \equiv$ *number of documents that mention str at least once.*

Generalized document frequency, $df_k(str)$, is the number of documents that mention str at least k times. For expository convenience, we use cdf_k for the cumulative document frequency: $cdf_k(str) \equiv \sum_{i=k}^{\infty} df_i(str)$. We can work directly with cdf_k as in [20], or alternatively, cdf_k can be used to compute more standard quantities such as frequency and df_k:

$$freq = cdf_1$$

$$df_k = cdf_k - cdf_{k+1}$$

Evaluation of the syntactic annotation in EPEC, the Reference Corpus for the Processing of Basque

Larraitz Uria, Ainara Estarrona, Izaskun Aldezabal, Maria Jesús Aranzabe,
Arantza Díaz de Ilarraza, Mikel Iruskieta

IXA Group (Natural Language Processing)
University of the Basque Country
Computer Science Faculty
Manuel Lardizabal Pasealekua 1
E-20018 Donostia

{larraitz.uria, ainara.estarrona, izaskun.aldezabal, maxux.aranzabe, a.diazdeilarraza, mikel.iruskieta}@ehu.es

Abstract. The aim of this work is to evaluate the dependency-based annotation of EPEC (the Reference Corpus for the Processing of Basque) by means of an experiment: two annotators have syntactically tagged a sample of the mentioned corpus in order to evaluate the agreement-rate between them and to identify those issues that have to be improved in the syntactic annotation process. In this article we present the quantitative and qualitative results of this evaluation.

Key Words: Basque corpus, dependency-based syntactic annotation, evaluation, annotators' agreement-rate, Kappa agreement index.

1. Introduction

This work has been carried out in the framework of the Ixa research group[1], where resources such as data-bases and corpora annotated at different linguistic levels are being developed.

The EPEC corpus [1], considered in the Ixa group a reference corpus for the processing of Basque, is so far annotated at syntactic level, with dependencies' relations; and a part of the semantic annotation (the nominal part) is also finished.

Every annotation process has to be evaluated in order to warranty its quality. In this paper, we present the qualitative and quantitative evaluation of the dependency-based annotation of a sample of EPEC. The aim of this evaluation is twofold: to measure the agreement-rate between the annotators and to identify those issues that have to be improved in the syntactic annotation process.

The paper is organized as follows: in section 2 we explain some features of the EPEC corpus. Section 3 deals with the model adopted for the syntactic analysis and

[1] http://ixa.si.ehu.es/Ixa

Reducing Noise in Labels and Features for a Real World Dataset: Application of NLP Corpus Annotation Methods

Rebecca J. Passonneau*, Cynthia Rudin†, Axinia Radeva‡, and Zhi An Liu§

Columbia University, New York, NY 10027, USA
*becky@cs.columbia.edu, (†cr2363|§zl2153)@columbia.edu,
‡axinia@hotmail.com

Abstract. This paper illustrates how a combination of information extraction, machine learning, and NLP corpus annotation practice was applied to a problem of ranking vulnerability of structures (service boxes, manholes) in the Manhattan electrical grid. By adapting NLP corpus annotation methods to the task of knowledge transfer from domain experts, we compensated for the lack of operational definitions of components of the model, such as *serious event.* The machine learning depended on the ticket classes, but it was not the end goal. Rather, our rule-based document classification determines both the labels of examples and their feature representations. Changes in our classification of events led to improvements in our model, as reflected in the AUC scores for the full ranked list of over 51K structures. The improvements for the very top of the ranked list, which is of most importance for prioritizing work on the electrical grid, affected one in every four or five structures.

1 Introduction

This paper illustrates how a combination of information extraction, machine learning, and NLP corpus annotation techniques was applied to a problem of ranking vulnerability of structures (manholes and service boxes) in the Manhattan electrical grid. Institutions of all sorts collect and archive large amounts of data. The value of the data to the institution depends in part on whether methods can be developed to make use of it. Information extraction, defined as the task of organizing and normalizing data taken from unstructured text in order to populate tables in structured databases, has obvious relevance, as does machine learning. Automated techniques for corpus annotation clearly support the task of information extraction. What is less obvious, and potentially of greater impact, is that the practices developed in the NLP community for manual annotation and classification of documents provide an effective means to arrive at clearer problem definitions. By adapting NLP corpus annotation methods to the task of knowledge transfer from domain experts, we compensated for the lack of operational definitions of components of the model, such as *serious event.*

During the 1970s, the Consolidated Edison Company instituted a program of Emergency Control System (ECS) tickets to document calls from customers about potential problems in the electrical grid. Beginning in 1986, after Hurrican Gloria, the ECS program expanded and became more fully utilized. As the

Unsupervised Classification of Verb Noun Multi-Word Expression Tokens

Mona T. Diab and Madhav Krishna
mdiab@ccls.columbia.edu, madhkrish@gmail.com

Columbia University, New York, NY 10115

Abstract. We address the problem of classifying multiword expression tokens in running text. We focus our study on Verb-Noun Constructions (VNC) that vary in their idiomaticity depending on context. VNC tokens are classified as either idiomatic or literal. Our approach hinges upon the assumption that a literal VNC will have more **in common** with its component words than an idiomatic one. Commonality is measured by contextual overlap. To this end, we set out to explore different contextual variations and different similarity measures. We also identify a new data set OPAQUE that comprises only non-decomposable VNC expressions. Our approach yields state of the art performance with an overall accuracy of 77.56% on a TEST data set and 81.66% on the newly characterized data set OPAQUE.

1 Introduction

A Multi-Word Expression (MWE), for our purposes, can be defined as a multi-word unit that refers to a single concept, for example - *kick the bucket*, *spill the beans*, *make a decision*, etc. An MWE typically has an idiosyncratic meaning that is *more* or *different* than the meaning of its component words. An MWE meaning is transparent, i.e. predictable, in as much as the component words in the expression relay the meaning portended by the speaker compositionally. Accordingly, MWEs vary in their degree of meaning compositionality; compositionality is correlated with the level of idiomaticity. An MWE is compositional if the meaning of an MWE as a unit can be predicted from the meaning of its component words such as in *make a decision* meaning *to decide.* If we conceive of idiomaticity as being a continuum, the more idiomatic an expression, the less transparent and the more non-compositional it is. Some MWEs are more predictable than others, for instance, *kick the bucket*, when used idiomatically to mean *to die*, has nothing in common with the literal meaning of either *kick* or *bucket*, however, *make a decision* is very clearly related to *to decide.* Both of these expressions are considered MWEs but have varying degrees of compositionality and predictability.

MWEs are pervasive in natural language, especially in the ever more abundant web based texts and speech genres. Identifying MWEs and understanding their meaning is essential to language understanding, hence they are of crucial

Semantic Mapping for Related Term Identification

Rafael E. Banchs[1]

[1] Barcelona Media Innovation Centre,
Av. Diagonal 177, Planta 9, 08018 Barcelona, Spain
rafael.banchs@barcelonamedia.org

Abstract. In this work, we explore the combined use of latent semantic analysis (LSA) and multidimensional scaling (MDS) for identifying related concepts and terms. We approach the problem of related term identification by constructing low-dimensional embeddings where related terms are clustered together, and such clusters are spatially arranged according to the semantic relationships among the terms they include. In this work, we demonstrate the proposed methodology for a specific part-of-speech (verbs) of the Spanish language, by using dictionary-based definitions. We also comment on the future use of this experimental framework in the context of other natural language processing tasks such as opinion mining, topic detection and automatic summarization.

Keywords: Vector Space Model, Latent Semantic Analysis, Multidimensional Scaling, Related Term Identification.

1 Introduction

The vector-space model has been extensively used in information retrieval and other text-mining applications. Within this particular context, some prominent techniques such as latent semantic analysis [4] and probabilistic latent semantic analysis [6] have been developed in order to obtain more efficient vector-space representations in terms of space dimensionality reduction and feature-based structural characterization. In this work, we attempt to combine latent semantic analysis along with multidimensional scaling with the objective of further reducing dimensional complexity while preserving structural characterization.

The methodology being proposed in this article is not intended to constitute a new solution for any specific text analysis problem, but rather to provide an experimental framework for exploring and evaluating text processing alternatives which could benefit from the tractability of very-low-dimensional data representations. In this work, the proposed methodology is presented and illustrated with the problem of related term identification [1]. An alternative approach for concept association representations by means VOS projections has been already proposed in [7].

The paper is structured as follows. First, in section 2, a brief overview of latent semantic indexing and multidimensional scaling is presented. Then, in section 3, the proposed methodology is described in detail. In section 4, the experimental work is

An Improved Automatic Term Recognition Method for Spanish

Alberto Barrón-Cedeño[1,2], Gerardo Sierra[1],
Patrick Drouin[3], and Sophia Ananiadou[4]

[1] Engineering Institute,
Universidad Nacional Autónoma de México, MEXICO
[2] Department of Information Systems and Computation,
Universidad Politécnica de Valencia, SPAIN
[3] Observatoire Linguistique Sense-Text,
Université de Montréal, CANADA
[4] National Centre for Text Mining, UK
alberto@pumas.ii.unam.mx, gsierram@ii.unam.mx,
patrick.drouin@umontreal.ca, sophia.ananiadou@manchester.ac.uk

Abstract. The *C-value/NC-value* algorithm, a hybrid approach to automatic term recognition, has been originally developed to extract multiword term candidates from specialised documents written in English. Here, we present three main modifications to this algorithm that affect how the obtained output is refined. The first modification aims to maximise the number of real terms in the list of candidates with a new approach for the stop-list application process. The second modification adapts the *C-value* calculation formula in order to consider single word terms. The third modification changes how the term candidates are grouped, exploiting a lemmatised version of the input corpus. Additionally, size of candidate's context window is variable. We also show the necessary linguistic modifications to apply this algorithm to the recognition of term candidates in Spanish.

1 Introduction

The *C-value/NC-value* algorithm [3] is the base of the Termine suite for automatic multiword terms recognition in specialised documents in English[1]. This tool, developed at the National Centre for Text Mining in Manchester, UK, was originally created for the extraction of biomedical terms. This algorithm has been applied to Automatic Term Recognition (ATR) over different languages such as English [3] and Japanese [10]. Additionally, it is the base for an algorithm designed for term extraction in Chinese [4]. A first essay has started to adapt it to handle documents in Spanish [1].

In this paper, we describe the improvements carried out over different stages of the algorithm. Additionally, we show the necessary adaptations for exploiting this algorithm on ATR of terms in Spanish texts. About the distribution of

[1] http://www.nactem.ac.uk/software/termine/

Bootstrapping a Verb Lexicon for Biomedical Information Extraction

Giulia Venturi[1], Simonetta Montemagni[1], Simone Marchi[1],
Yutaka Sasaki[2,3], Paul Thompson[2,3], John McNaught[2,3] and Sophia Ananiadou[2,3]

[1] Istituto di Linguistica Computazionale, CNR, Pisa, Italy
[2] School of Computer Science, University of Manchester, UK
[3] National Centre for Text Mining, University of Manchester, UK
{giulia.venturi, simonetta.montemagni, simone.marchi}@ilc.cnr.it
{yutaka.sasaki, paul.thompson, jock.mcnaught, sophia.ananiadou}@manchester.ac.uk

Abstract. The extraction of information from texts requires resources that contain both syntactic and semantic properties of lexical units. As the use of language in specialized domains, such as biology, can be very different to the general domain, there is a need for domain-specific resources to ensure that the information extracted is as accurate as possible. We are building a large-scale lexical resource for the biology domain, providing information about predicate-argument structure that has been bootstrapped from a biomedical corpus on the subject of E. Coli. The lexicon is currently focussed on verbs, and includes both automatically-extracted syntactic subcategorization frames, as well as semantic event frames that are based on annotation by domain experts. In addition, the lexicon contains manually-added explicit links between semantic and syntactic slots in corresponding frames. To our knowledge, this lexicon currently represents a unique resource within in the biomedical domain.

Keywords: domain-specific lexical resources, lexical acquisition, syntax-semantics linking, Information Extraction, Biological Language Processing

1 Introduction

It is well known that Information Extraction applications require sophisticated lexical resources to support their processing goals. In particular, accurate applications focussed on extraction of event information from texts require resources containing both syntactic and semantic information. Many applications could benefit from lexical resources providing an exhaustive account of the semantic and syntactic combinatorial properties of lexical units conveying event information.

The need for such resources increases when dealing with texts belonging to a specialized domain such as biology. There are several reasons for requiring domain-specific lexical resources. Even more than in general language, within specialized domains, much lexical knowledge is idiosyncratically related to the individual behavior of lexical units. In particular, it can be the case that the types of events mentioned

TermeX: A Tool for Collocation Extraction

Davor Delač, Zoran Krleža, Jan Šnajder,
Bojana Dalbelo Bašić, and Frane Šarić

Faculty of Electrical Engineering and Computing
University of Zagreb, Croatia
{davor.delac, zoran.krleza, jan.snajder, bojana.dalbelo, frane.saric}@fer.hr

Abstract. Collocations – word combinations occurring together more often than by chance – have a wide range of NLP applications. Many approaches for automating collocation extraction based on lexical association measures have been proposed in the literature. This paper presents *TermeX* – a tool for efficient extraction of collocations based on a variety of association measures. *TermeX* implements POS filtering and lemmatization, and is capable of extracting collocations up to length four. We address trade-offs between high memory consumption and processing speed and propose an efficient implementation. Our implementation allows for processing time linear to corpus size and memory consumption linear to the number of word types.

1 Introduction

Collocations are a lexical phenomenon that has a linguistic and lexicographic status. In [1] collocations are defined as "institutionalized phrases", whereas [2] defines them as "word combinations occurring together more often than by chance." There is a wide range of possible applications for collocation extraction in NLP such as word sense disambiguation [3, 4], natural language generation [5], and machine translation [6]. However, collocation extraction is a time consuming task for a human and requires the expertise of a professional lexicographer. Therefore, many approaches for automating collocation extraction have been proposed in the literature.

This paper presents a collocation extraction tool called *TermeX*. This tool is meant for construction of terminology lexica with possible applications in NLP. *TermeX* focuses on the use of lexical association measures (AMs) to automatically extract collocations up to length four (4-grams). It provides the user with a variety of association measures to choose from as well as the ability to manually select valid collocations from those extracted automatically, allowing for construction of domain-specific terminology lexica. Besides English, *TermeX* currently supports Croatian language. In order to improve collocation extraction, *TermeX* implements POS filtering and lemmatization, the latter being important due to morphological complexity of Croatian language. This paper also covers some of the implementation issues such as trade-offs between high memory consumption and processing speed. *TermeX* is able to cope with large amounts

Guessers for Finite-State Transducer Lexicons

Krister Lindén

Department of General Linguistics, P.O. Box 9, FIN-00014 University of Helsinki
Krister.Linden@Helsinki.fi

Abstract. Language software applications encounter new words, e.g., acronyms, technical terminology, names or compounds of such words. In order to add new words to a lexicon, we need to indicate their inflectional paradigm. We present a new generally applicable method for creating an entry generator, i.e. a paradigm guesser, for finite-state transducer lexicons. As a guesser tends to produce numerous suggestions, it is important that the correct suggestions be among the first few candidates. We prove some formal properties of the method and evaluate it on Finnish, English and Swedish full-scale transducer lexicons. We use the open-source *Helsinki Finite-State Technology* [1] to create finite-state transducer lexicons from existing lexical resources and automatically derive guessers for unknown words. The method has a recall of 82-87 % and a precision of 71-76 % for the three test languages. The model needs no external corpus and can therefore serve as a baseline.

1 Introduction

New words and new usages of old words are constantly finding their way into daily language use. This is particularly prominent in rapidly developing domains such as biomedicine and technology. The new words are typically acronyms, technical terminology, loan words, names or compounds of such words. They are likely to be unknown by most hand-made morphological analyzers. In some applications, hand-made guessers are used for covering the low-frequency vocabulary or the strings are simply added as such.

Mikheev [2] and [16] noted that words unknown to the lexicon present a substantial problem to part-of-speech tagging and he presented a very effective supervised method for inducing a guesser from a lexicon and an independent training corpus. Oflazer & al. [3] presented an interactive method for learning morphologies and pointed out that an important issue in the wholesale acquisition of open-class items is that of determining which paradigm a given citation form belongs to.

Recently, unsupervised acquisition of morphologies from scratch has been studied as a general problem of morphology induction in order to automate the morphology building procedure. For overviews, see Wicentowski [4] and Goldsmith [5]. If we do not need a full analysis, but only wish to segment the words into morph-like units, we can use segmentation methods like Morfessor [6]. For a comparison of some recent successful segmentation methods, see the Morpho Challenge [7].

Although unsupervised methods have advantages for less-studied languages, for the well-established languages, we have access to fair amounts of lexical training ma-

Combining Language Modeling and Discriminative Classification for Word Segmentation

Dekang Lin

Google, Inc.
1600 Amphitheater Parkway, Mountain View, CA, USA, 94043
lindek@google.com

Abstract. Generative language modeling and discriminative classification are two main techniques for Chinese word segmentation. Most previous methods have adopted one of the techniques. We present a hybrid model that combines the disambiguation power of language modeling and the ability of discriminative classifiers to deal with out-of-vocabulary words. We show that the combined model achieves 9% error reduction over the discriminative classifier alone.

Keywords: Segmentation, Maximum Entropy, Language Model

1 Introduction

The problem of word segmentation is to identify word boundaries in languages, such as Chinese, Japanese and Thai, where such boundaries are not explicitly marked with white spaces. Word segmentation is the first step in processing the text in these languages and its quality often affects all downstream components.

The main two challenges for word segmentation are the resolution of ambiguities and the handling of out-of-vocabulary (OOV) words.

Ambiguity is pervasive in word segmentation. Consider the following fragment of a Chinese sentence:

…中国家鼓励… (in … the country encourages …) (1)

The correct segmentation is:

中	国家	鼓励
in	country	encourages

However, another (incorrect) candidate is:

中国	家	鼓励
China	home	encourages

An obvious way to resolve the ambiguities is through language modeling. The best segmentation of an input sentence $C_1^n = C_1C_2...C_n$ is the one where the resulting

Formal Grammar for Hispanic Named Entities Analysis

Grettel Barceló, Eduardo Cendejas, Grigori Sidorov, and Igor A. Bolshakov

Centro de Investigación en Computación
Instituto Politécnico Nacional
Mexico City, Mexico
{gbarceloa07, ecendejasa07}@sagitario.cic.ipn.mx
{sidorov, igor}@cic.ipn.mx

Abstract. A task that has been widely studied in the field of natural language processing is the Named Entity Recognition (NER). A great number of approaches have been developed to deal with the identification and classification of named entity strings in specific- and open-domains. Nevertheless, external modules have to be incorporated into many of the NER systems in order to solve the interpretation problems derived from proper nouns. In this article our focus will be on the study of ambiguity in Hispanic Nominal Sequences which constitution assumes three main problems: (1) the association of given names and/or surnames; (2) the composition of such elements by means of a connector; (3) and the duality of given name/surname. In order to analyze the magnitude of the problem, two gazetteers were made, one with 93998 given names and the other with 13779 surnames. The gazetteers entries were used as terminal symbols of the proposed grammar to determine the valid interpretations in the nominal sequences; this is done by means of an automatic labeling of all the elements the nominal sequences are made of.

1 Introduction

Named entity recognition deals with the identification and the classification of strings that belong to proper nouns. Several categories have been established for the classification process; some of the most important categories are: person, organization and location. NER systems can be used independently [1] or as modules of text pre-processing in different NLP applications where the recognition efficacy impacts directly on its quality. These systems have been put to work in machine translation tasks [2], [3], information extraction [4], [5], [6], information retrieval [7], [8] and question answering [9], [10].

Based on the text analysis of several genres, studies have proved that there is a high percentage in occurrence of proper nouns. Since the origin of NER in 1995, and up to today, several heuristics and algorithms have been proposed, with hand crafted rules or statistical approaches.

However, ambiguity problems are still persistent in the identification as well as in the classification stage. Three main errors are mentioned in [11] that can be found in some categories, such as:

Automatic Extraction of Clause Relationships from a Treebank

Oldřich Krůza and Vladislav Kuboň

Charles University in Prague
Institute of Formal and Applied Linguistics
Malostranské nám. 25, Prague
Czech Republic
kruza@ufal.mff.cuni.cz, vk@ufal.mff.cuni.cz

Abstract. The paper concentrates on deriving non-obvious information about clause structure of complex sentences from the Prague Dependency Treebank. Individual clauses and their mutual relationship are not explicitly annotated in the treebank, therefore it was necessary to develop an automatic method transforming the original annotation concentrating on the syntactic role of individual word forms into a scheme describing the relationship between individual clauses. The task is complicated by a certain degree of inconsistency in original annotation with regard to clauses and their structure. The paper describes the method of deriving clause-related information from the existing annotation and its evaluation.

1 Introduction

One of the major factors which changed linguistics during the past twenty years was without a doubt a strong stress on building and exploiting large annotated corpora of natural languages. They serve nowadays as a primary source of evidence for the development and evaluation of linguistic theories and applications.

Although the corpora are extremely important source of data, they are not omnipotent. The more elaborated annotation scheme the authors use, the more problems with linguistic phenomena they have to solve. Creating a consistently annotated treebank requires making a large number of decisions about a particular annotation of particular linguistic phenomena. The more elaborated and detailed is the annotation, the easier it is to find phenomena which are annotated in a seemingly inconsistent way. If the annotation is really well-designed and consistent, then it should be possible to extract an information hidden in the corpus or treebank even in case that a particular phenomenon we are interested in was not annotated explicitly.

This paper describes an attempt to do precisely that - to extract an information which may be useful for research of a particular linguistic phenomenon from the treebank, where this phenomenon is not explicitly tagged. The axtracted information should provide a linguistic basis for research in the fields of natural language parsing and information retrieval (exploiting linguistically motivated

A general method for transforming standard parsers into error-repair parsers*

Carlos Gómez-Rodríguez[1], Miguel A. Alonso[1], and Manuel Vilares[2]

[1] Departamento de Computación, Universidade da Coruña (Spain)
{cgomezr, alonso}@udc.es
[2] Escuela Superior de Ingeniería Informática, Universidade de Vigo (Spain)
vilares@uvigo.es

Abstract. A desirable property for any system dealing with unrestricted natural language text is robustness, the ability to analyze any input regardless of its grammaticality. In this paper we present a novel, general transformation technique to automatically obtain robust, error-repair parsers from standard non-robust parsers. The resulting error-repair parsing schema is guaranteed to be correct when our method is applied to a correct parsing schema verifying certain conditions that are weak enough to be fulfilled by a wide variety of parsers used in natural language processing.

1 Introduction

In real-life domains, it is common to find natural language sentences that cannot be parsed by grammar-driven parsers, due to insufficient coverage (the input is well-formed, but the grammar cannot recognize it) or ill-formedness of the input (errors in the sentence or errors caused by input methods). A standard parser will fail to return an analysis in these cases. A *robust parser* is one that can provide useful results for such extragrammatical sentences.

The methods that have been proposed to achieve robustness in parsing fall mainly into two broad categories: those that try to parse well-formed fragments of the input when a parse for the complete sentence cannot be found (partial parsers, such as that described in [6]) and those which try to assign a complete parse to the input sentence by relaxing grammatical constraints, such as *error-repair parsers*, which can find a complete parse tree for sentences not covered by the grammar by supposing that ungrammatical strings are corrupted versions of valid strings.

The problem of repairing and recovering from syntax errors during parsing has received much attention in the past (see for example the list of references provided in the annotated bibliography of [5, section 18.2.7]) and recent years (see for example [15, 17, 2, 7, 1, 11]). In this paper, we try to fill the gap between standard and error-repair parsing by proposing a transformation for automatically obtaining error-repair parsers, in the form of *error-repair parsing schemata*, from standard parsers defined as *parsing schemata*.[3]

* Partially supported by Ministerio de Educación y Ciencia and FEDER (HUM2007-66607-C04) and Xunta de Galicia (PGIDIT07SIN005206PR, INCITE08E1R104022ES, INCITE08ENA305025ES, INCITE08PXIB302179PR and Rede Galega de Procesamento da Linguaxe e Recuperación de Información)

[3] *Schemata* is the plural form of the singular noun *schema*.

Topic-focus articulation from the semantic point of view

Marie Duží

VSB-Technical University Ostrava, Czech Republic
marie.duzi@vsb.cz

Abstract. In the paper we show that sentences differing only in topic-focus articulation have different logical structures, and thus they also have different truth-conditions. Our analysis is based on the procedural semantics of Transparent Intensional Logic (TIL) assigning to sentences hyperpropositions as their structured meanings. We analyse the phenomena of presupposition connected with a topic and allegation triggered by a focus of a sentence so that relevant consequences can be formally derived.

1 Introduction

In the invited talk [4], presented at CICLing 2008, Eva Hajičová argued that the problem of topic-focus articulation is a semantic, rather than pragmatic problem. We agree, and to put her arguments still on a more solid ground, we are going to demonstrate the semantic nature of the topic-focus difference by its *logical* analysis. To this end we apply procedural semantics of Transparent Intensional Logic (TIL) and assign (algorithmically structured) procedures to expressions as their meaning. As a result, we furnish sentences differing only in the topic-focus articulation with different structured meanings producing different propositions.

As sample sentences we analyse (slightly modified) examples adduced in [4]. Moreover, we present general schemata of a logical structure of arguments according whether the phenomena of presupposition or allegation are the case. These relations are defined in [4], where Hajičová shows that the clause standing in the topic often induces the case of presupposition, whereas a focus-clause is connected with allegation. If a presupposition *Q* of a given proposition *P* is not true, then *P* as well as negated *P* have no truth-value. In other words, *Q* is entailed both by *P* and non-*P*. On the other hand, if *Q* is an allegation of *P*, then *P* entails, but does not presuppose, *Q*. If non-*P* is the case, we cannot deduce anything about the truth of *Q*. Since our logic is a hyper-intensional logic of ***partial functions***, we analyse sentences with presuppositions in a natural way. We furnish them with hyper-propositions that produce propositions with truth-value gaps. Having a rigorous, fine-grained analysis at our disposal, we can easily infer the relevant consequences. Thus our logic meets the philosophical and linguistic desiderata formulated in [4].

The paper is organised as follows. After briefly introducing TIL philosophy and its basic notions in Section 2, the main Section 3 describes the method of analysing sentences with topic-focus articulation. Concluding Section 4 presents the direction of future research and a few notes on TIL implementation via the *TIL-Script* functional programming language.

The Value of Weights in Automatically Generated Text Structures

Dana Dannélls

NLP Research Unit, Department of Swedish Language,
University of Gothenburg, Sweden
dana.dannells@svenska.gu.se

Abstract. One question that arises if we want to evolve generation techniques to accommodate Web ontologies is how to capture and expose the relevant ontology content to the user. This paper presents an attempt to answer the question about how to select the ontology statements that are significant for the user and present those statements in a way that helps the user to learn. Our generation approach combines bottom-up and top-down techniques with enhanced comparison methods to tailor descriptions about a concept described in an ontology. A preliminary evaluation indicates that the process of computing preferable property weights in addition to enhanced generation methods has a positive effect on the text structure and its content. Future work aims to assign grammar rules and lexical entries in order to produce coherent texts that follow on from the generated text structures in several languages.

Key words: NLG, Ontology, Semantic Web.

1 Introduction

The ability to generate natural language text from web ontology languages and more generally knowledge bases that are encoded in RDF (Resource Description Framework) imposes new demands on natural language generators that aim to produce written text either for textual presentation or for eventual use by text-to-speech system. One of these demands concerns the process of text planning. Text planning, also referred to *Document Planning* [20], is the process responsible for producing a specification of the text's content and structure. The fact that aspects such as the user characteristics, e.g., cognitive state, desires, the background domain knowledge, and linguistic properties must be taken into account and computed simultaneously during planning makes this process computationally hard and so far there has been little success in computing a general model with a suitable structure for generating from ontologies in general and from web ontologies in particular. This brings a need to find alternative strategies to generate knowledge from ontology languages, or alternatively to adapt previously presented ideas to the new emerging technology standards.

Recent attempts to develop natural language generators that support the Web Ontology Language (OWL) and similar Semantic Web languages,[1] treat the

[1] http://www.w3.org/TR/

AORTE for Recognizing Textual Entailment

Reda Siblini and Leila Kosseim

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

Abstract. In this paper we present the use of the AORTE system in recognizing textual entailment. AORTE allows the automatic acquisition and alignment of ontologies from text. The information resulted from aligning ontologies created from text fragments is used in classifying textual entailment. We further introduce the set of features used in classifying textual entailment. At the TAC RTE4 challenge the system evaluation yielded an accuracy of 68% on the two-way task, and 61% on the three way task using a simple decision tree classifier.

1 Introduction

In this paper we present a novel method of recognizing textual entailment. Textual entailment is defined as *"a relationship between a coherent text T and a language expression, which is considered as a hypothesis, H. We say that T entails H (H is a consequent of T), if the meaning of H, as interpreted in the context of T, can be inferred from the meaning of T"* [1].
For example, the text:
(T): *Jurassic Park is a novel written by Michael Crichton.*
Entails the following hypothesis (among others):
(H1): *Michael Crichton is an author.*
(H2): *Jurassic Park is a book.*
(H3): *Michael Crichton is the author of the book Jurassic Park.*

Recognizing textual entailment is a fundamental task to many applications in natural language processing, such as in *Information Retrieval* where retrieving relevant documents could be seen as finding documents containing the text that entails the information we are looking for, in *Information Extraction* where the extraction of information is based on a set of templates that entail the information that we would like to extract, in *Question Answering* where candidate answers are snippets that entail the question we want to answer, and in *Summarization* where redundancy can be avoided by detecting textual entailment.

The remainder of the paper is organized as follows. Section 2 presents related work in recognizing textual entailment. In Section 3 we give an overview of our

Semi-supervised Word Sense Disambiguation using the Web as Corpus

Rafael Guzmán-Cabrera[1,2], Paolo Rosso[2], Manuel Montes-y-Gómez[3],
Luis Villaseñor-Pineda[3], David Pinto-Avendaño[4]

[1] FIMEE, Universidad de Guanajuato, Mexico
guzmanc@salamanca.ugto.mx
[2] NLE Lab, DSIC, Universidad Politécnica de Valencia, Spain
prosso@dsic.upv.es
[3] LabTL, Instituto Nacional de Astrofísica, Óptica y Electrónica, Mexico
{mmontesg, villasen}@inaoep.mx
[4] FCC, Benemérita Universidad Autónoma de Puebla, Mexico
dpinto@cs.buap.mx

Abstract. As any other classification task, Word Sense Disambiguation requires a large number of training examples. These examples, which are easily obtained for most of the tasks, are particularly difficult to obtain for this case. Based on this fact, in this paper we investigate the possibility of using a Web-based approach for determining the correct sense of an ambiguous word based only in its surrounding context. In particular, we propose a semi-supervised method that is specially suited to work with just a few training examples. The method considers the automatic extraction of unlabeled examples from the Web and their iterative integration into the training data set. The experimental results, obtained over a subset of ten nouns from the SemEval lexical sample task, are encouraging. They showed that it is possible to improve the baseline accuracy of classifiers such as Naïve Bayes and SVM using some unlabeled examples extracted from the Web.

1 Introduction

It is well known that, in all languages, some words may have several different meanings or senses. For example, in English, the word “bank” can either mean a financial institution or a sloping raised land. Related to this language phenomenon, the task of *Word Sense Disambiguation* (WSD) considers the assignment of the correct sense to such ambiguous words based on their surrounding context [6].

There are two main kinds of methods to carry out the task of WSD. On the one hand, the knowledge-based methods, which disambiguate words by comparing their context against information from a predefined lexical resource such as Wordnet [1, 3]. On the other hand, *corpus-based methods,* which achieve the sense disambiguation by applying rules that were automatically learned from a sense tagged corpus [14]. Recent reports [8] indicate that corpus-based methods tend to be more precise than knowledge-based ones. Nevertheless, due to the lack of large sense tagged cor-

Semi-supervised Clustering for Word Instances and Its Effect on Word Sense Disambiguation

Kazunari Sugiyama and Manabu Okumura

Precision and Intelligence Laboratory, Tokyo Institute of Technology,
4259 Nagatsuta, Midori, Yokohama, Kanagawa 226-8503, Japan
sugiyama@lr.pi.titech.ac.jp, oku@pi.titech.ac.jp

Abstract. We propose a supervised word sense disambiguation (WSD) system that uses features obtained from clustering results of word instances. Our approach is novel in that we employ semi-supervised clustering that controls the fluctuation of the centroid of a cluster, and we select seed instances by considering the frequency distribution of word senses and exclude outliers when we introduce "must-link" constraints between seed instances. In addition, we improve the supervised WSD accuracy by using features computed from word instances in clusters generated by the semi-supervised clustering. Experimental results show that these features are effective in improving WSD accuracy.

1 Introduction

Many words have multiple meanings depending on the context in which they are used. For example, among the possible senses of the verb "run" are "to move fast by using one's feet" and "to direct or control." Word sense disambiguation (WSD) is the task of determining the meaning of such an ambiguous word in its context. In this paper, we apply semi-supervised clustering by introducing sense-tagged instances (we refer to them as "seed instances" in the following) to the supervised WSD process. Our approach is based on the following intuitions: (1) in the case of word instances, we can use sense-tagged word instances from various sources as supervised instances, and (2) the features computed from word instances in clusters generated by our semi-supervised clustering are effective in supervised WSD since word instances clustered around sense-tagged instances may have the same sense. Existing semi-supervised clustering approaches solely focus on introducing constraints and learning distances and overlook control of the fluctuation of the cluster's centroid. In addition, to enable highly accurate semi-supervised clustering, it is important to consider how to select seed instances and how to introduce constraints between the seed instances. Regarding seed instances, we have to pay attention to the frequency distribution of word senses when selecting seed instances as well as the way of introducing "must-link" constraints, since outlier instances may exist when we select seed instances with the same sense.

In this paper, we describe our semi-supervised clustering approach that controls the fluctuation of the centroid of a cluster and propose a way of introducing appropriate seed instances and constraints. In addition, we explain our WSD approach using features computed from word instances that belong to clusters generated by the semi-supervised clustering. Our approach is novel in that we employ semi-supervised clustering that controls the fluctuation of the centroid of a cluster, and we select seed instances by considering the frequency distribution of word senses and exclude outliers when we introduce "must-link" constraints between seed instances.

2 Related Work

2.1 Semi-supervised Clustering

The semi-supervised clustering methods can be classified into *constraint-based* and *distance-based*. Constraint-based methods rely on user-provided labels or constraints to guide the algorithm toward a more appropriate data partitioning. For example, Wagstaff et al. [12, 13] introduced two types of constraint – "must-link" (two instances have to

Alleviating the Problem of Wrong Coreferences in Web Person Search

Octavian Popescu, Bernardo Magnini
papsi@racai.ro, magnini@fbk.eu

Abstract. In this paper we present a system for the Web People Search task, which is the task of clustering together the pages referring to the same person. The vector space model approached is modified in order to develop a more flexible clustering technique. We have implemented a dynamic weighting procedure for the attributes common to different cluster in order to maximize the between cluster variance with respect with the within cluster variance. We show that in this way the undesired collateral effect such as superposition and masking are alleviated. The system we present obtains similar results to the ones reported by the top three systems presented at the SEMEVAL 2007 competition.

Keywords: Web People Search, Cascade Clustering, Dynamic Threshold, Masking, Superposition.

1 Introduction

The exponential development of the Web brings with it the need for Web search tools. While for a certain class of queries the search engines on the market offer good answers, there is a significant part of queries that remains unfulfilled. Consulting the first 200 results returned by Google for the query "Bush engineer", one can notice that those pages refer only to two different persons. And one can learn unexpected connections between the US president, "George W. Bush", and the word "Engineer". However, it will probably be more useful if a list with different persons named "Bush" who are engineers would be returned instead.

The task of clustering the result pages of a search engine for a person name query according to the person they refer to has been recently undertaken in the Web People Search competition (WPS), under the Semeval 2007 workshop ([1]). The WPS task is slightly different from the Cross-Document Coreference task ([5]), which is the task of establishing coreferences among the entities present in a corpus. This is because some pages may be clustered together even without indicating which person mentions are actually corefered. In this paper we present a system for WPS task and its performances on the WPS test and training data set. The technique is based on corefered entities, so it is applicable to the cross-document coreference task as well.

Improved Unsupervised Name Discrimination with Very Wide Bigrams and Automatic Cluster Stopping

Ted Pedersen

University of Minnesota, Duluth, MN 55812, USA

Abstract. We cast name discrimination as a problem in clustering short contexts. Each occurrence of an ambiguous name is treated independently, and represented using second–order context vectors. We calibrate our approach using a manually annotated collection of five ambiguous names from the Web, and then apply the learned parameter settings to three held-out sets of pseudo-name data that have been reported on in previous publications. We find that significant improvements in the accuracy of name discrimination can be achieved by using very wide bigrams, which are ordered pairs of words with up to 48 intervening words between them. We also show that recent developments in automatic cluster stopping can be used to predict the number of underlying identities without any significant loss of accuracy as compared to previous approaches which have set these values manually.

1 Introduction

Person name ambiguity is an increasingly common problem as more and more people have online presences via Web pages, social network sites, and blogs. Since many distinct people share the same or similar names, it is often difficult to sort through results returned by search engines and other tools when looking for information about a particular person. There are many examples of identity confusion that have been widely publicized. For example, television talk show host Charlie Rose included his friend George Butler the filmmaker in his list of notable deaths from 2008. The only problem was that this George Butler was still alive, and it was George Butler the recording company executive who had died.

In general the goal of name discrimination is to associate or group the occurrences of person names with their true underlying identities. Our approach is completely unsupervised, and relies purely on the written contexts surrounding the ambiguous name. Our goal is to group these contexts into some (unspecified) number of clusters, where each cluster is associated with a unique individual. We assume that named entity recognition (NER) has already been carried out, so the input consists of text where the occurrences of person names are already identified.

There are various ways to formulate solutions to the problems surrounding name ambiguity or identity confusion, and so this paper tries to clarify exactly

Enriching Statistical Translation Models using a Domain-independent Multilingual Lexical Knowledge Base

Miguel García, Jesús Giménez and Lluís Màrquez

TALP Research Center, LSI Department
Universitat Politècnica de Catalunya
Jordi Girona Salgado 1–3, E-08034, Barcelona
{tgarcia,jgimenez,lluism}@lsi.upc.edu

Abstract. This paper presents a method for improving phrase-based Statistical Machine Translation systems by enriching the original translation model with information derived from a multilingual lexical knowledge base. The method proposed exploits the Multilingual Central Repository (a group of linked WordNets from different languages), as a domain-independent knowledge database, to provide translation models with new possible translations for a large set of lexical tokens. Translation probabilities for these tokens are estimated using a set of simple heuristics based on WordNet topology and local context. During decoding, these probabilities are softly integrated so they can interact with other statistical models. We have applied this type of domain-independent translation modeling to several translation tasks obtaining a moderate but significant improvement in translation quality consistently according to a number of standard automatic evaluation metrics. This improvement is especially remarkable when we move to a very different domain, such as the translation of Biblical texts.

1 Introduction

One of the main criticisms against empirical methods in general, and Statistical Machine Translation (SMT) in particular, is their strong domain dependence. Since parameters are estimated from a corpus in a specific domain, the performance of the system on a different domain is often much worse. This flaw of statistical and machine learning approaches is well known and has been largely described in the NLP literature, for a variety of tasks, e.g., parsing [1], word sense disambiguation [2], and semantic role labeling [3].

In the case of SMT, domain dependence has very negative effects in translation quality. For instance, in the 2007 edition of the ACL MT workshop (WMT07), an extensive comparative study between in-domain and out-of-domain performance of MT systems built for several European languages was conducted [4]. Results showed a significant difference in MT quality between the two domains for all statistical systems, consistently according to a number of automatic evaluation metrics. In contrast, the differences reported in the case of rule-based or hybrid MT systems were less significant or inexistent, and even in some cases the performance of such systems out of the domain was higher than in the corpus domain. The reason is that, while these systems are

Exploiting Parallel Treebanks to Improve Phrase-Based Statistical Machine Translation

John Tinsley, Mary Hearne, and Andy Way

National Centre for Language Technology
Dublin City University, Ireland
{*jtinsley, mhearne, away*}*@computing.dcu.ie*

Abstract. Given much recent discussion and the shift in focus of the field, it is becoming apparent that the incorporation of syntax is the way forward for the current state-of-the-art in machine translation (MT). Parallel treebanks are a relatively recent innovation and appear to be ideal candidates for MT training material. However, until recently there has been no other means to build them than by hand. In this paper, we describe how we make use of new tools to automatically build a large parallel treebank and extract a set of linguistically motivated phrase pairs from it. We show that adding these phrase pairs to the translation model of a baseline phrase-based statistical MT (PBSMT) system leads to significant improvements in translation quality. We describe further experiments on incorporating parallel treebank information into PBSMT, such as word alignments. We investigate the conditions under which the incorporation of parallel treebank data performs optimally. Finally, we discuss the potential of parallel treebanks in other paradigms of MT.

1 Introduction

The majority of research in recent years in machine translation (MT) has centred around the phrase-based statistical approach. This paradigm involves translating by training models which make use of sequences of words, so-called phrase pairs, as the core translation model of the system [1]. These phrase pairs are extracted from aligned sentence pairs using heuristics over a statistical word alignment. While phrase-based models have achieved state-of-the-art translation quality, evidence suggests there is a limit as to what can be accomplished using only simple phrases, for example, satisfactory capturing of context-sensitive reordering phenomena between language pairs [2]. This assertion has been acknowledged within the field as illustrated by the recent shift in focus towards more linguistically motivated models.

Aside from the development of fully syntax-based models of MT, [3–6] to list a few, there have been many extensions and improvements to the phrase-based model which have endeavoured to incorporate linguistic information into the translation process. Examples of these can be seen in the work of [7] and [8] who make use of syntactic supertags and morphological information respectively. [9, 10] describes a phrase-based model which makes use of generalised templates while [11] exploit semantic information in the form of phrase-sense disambiguation. All of these approaches have a

Cross-Language Frame Semantics Transfer in Bilingual Corpora

R. Basili[1], D. De Cao[1], D. Croce[1], B. Coppola[2], A. Moschitti[2]

[1] Dept. of Computer Science,
University of Roma Tor Vergata, Roma, Italy
{basili,croce,decao}@info.uniroma2.it
[2] University of Trento, Italy
{coppola, moschitti}@disi.unitn.it

Abstract. Recent work on the transfer of semantic information across languages has been recently applied to the development of resources annotated with Frame information for different non-English European languages. These works are based on the assumption that parallel corpora annotated for English can be used to transfer the semantic information to the other target languages. In this paper, a robust method based on a statistical machine translation step augmented with simple rule-based post-processing is presented. It alleviates problems related to preprocessing errors and the complex optimization required by syntax-dependent models of the cross-lingual mapping. Different alignment strategies are here investigated against the Europarl corpus. Results suggest that the quality of the derived annotations is surprisingly good and well suited for training semantic role labeling systems.

1 Motivation

The availability of large scale semantic lexicons, such as Framenet ([1]), has allowed the adoption of a vaste family of learning paradigms in the automation of semantic parsing. Building on the so called *frame* semantic model, the Berkeley FrameNet project [1] has developed a frame-semantic lexicon for the core vocabulary of English since 1997. As defined in [2], a frame is a conceptual structure modeling a prototypical situation. A frame is evoked in texts through the occurrence of its lexical units (LU), i.e. predicate words (verbs, nouns, or adjectives) that linguistically expresses the situation of the frame. Each frame also specifies the participants and properties of the situation it describes, the so called frame elements (FEs), that are the Frame Semantics instantiation of semantic roles. For example the frame CATEGORIZATION has lexical units such as: *categorize,classify,classification,regard*. Semantic roles shared by these predicates, are the COGNIZER (i.e. the person who performs the categorization act), the ITEM construed or treated, the CATEGORY (i.e. the class which the item is considered a member of) and CRITERIA. Semantic Role Labeling (SRL) is the task of automatic labeling individual predicates togheter with their major roles (i.e. frame elements) as they are grammatically realized in input sentences. It has been a popular task since the availability of the PropBank and Framenet annotated corpora [3], the seminal work of [4] and the successful CoNLL evaluation campaigns [5]. Statistical machine learning methods, ranging from joint probabilistic models to support vector machines, have been largely adopted to provide accurate labeling, although inherently dependent on the availability of large scale annotated resources.

It has been observed that the so called resulting *resource scarcity problem* affects a large number of languages for which such annotated corpora are not available [6]. Recent works thus explored

A Parallel Corpus Labeled using Open and Restricted Domain Ontologies [*]

E. Boldrini, S. Ferrández, R. Izquierdo, D. Tomás, and J.L. Vicedo

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

Abstract. The analysis and creation of annotated corpus is fundamental for implementing natural language processing solutions based on machine learning. In this paper we present a parallel corpus of 4500 questions in Spanish and English on the touristic domain, obtained from real users. With the aim of training a question answering system, the questions were labeled with the expected answer type, according to two different ontologies. The first one is an open domain ontology based on Sekine's Extended Named Entity Hierarchy, while the second one is a restricted domain ontology, specific for the touristic field. Due to the use of two ontologies with different characteristics, we had to solve many problematic cases and adjusted our annotation thinking on the characteristics of each one. We present the analysis of the domain coverage of these ontologies and the results of the inter-annotator agreement. Finally we use a question classification system to evaluate the labeling of the corpus.

1 Introduction

A corpus is a collection of written or transcribed texts created or selected using clearly defined criteria. It is a selection of natural language texts that are representative of the state of the language or of a special variety of it. Corpus annotation is a difficult task due to the ambiguities of natural language. As a consequence, annotation is time-consuming, but is extremely useful for natural language processing tasks based on machine learning, such as word sense disambiguation, named entity recognition or parsing.

Question answering (QA) is the task that, given a collection of documents (that can be a local collection or the World Wide Web), retrieves the answers to queries in natural language. The purpose of this work is the development of a corpus for training a question classification system. Question classification is one of the tasks carried out in a QA system. It assigns a class or category to the

[*] This research has been partially funded by the Spanish Government under project CICyT number TIC2003-07158-C04-01 and by the European Commission under FP6 project QALL-ME number 033860.

Language Identification on the Web: Extending the Dictionary Method

Radim Řehůřek[1] and Milan Kolkus[2]

[1] Masaryk University in Brno, `xrehurek@fi.muni.cz`
[2] Seznam.cz, a.s., `milan.kolkus@firma.seznam.cz`

Abstract. Automated language identification of written text is a well-established research domain that has received considerable attention in the past. By now, efficient and effective algorithms based on character n-grams are in use, mainly with identification based on Markov models or on character n-gram profiles. In this paper we investigate the limitations of these approaches when applied to real-world web pages. The challenges to be overcome include language identification on very short texts, correctly handling texts of unknown language and texts comprised of multiple languages. We propose and evaluate a new method, which constructs language models based on word relevance and addresses these limitations. We also extend our method to allow us to efficiently and automatically segment the input text into blocks of individual languages, in case of multiple-language documents.

1 Motivation

The amount of information available on the net is staggering and still growing at a fast pace. To make this information available, applications have sprung up to fill the void and gather, process and present Web information to the knowledge-hungry user. Unfortunatelly, documents on the Web have historically been created with human reader in mind, in formats such as HTML, and are not readily understandable by computers. Although XML and semantic markup (e.g. the `xml:lang` attribute, or the `<div lang="en">` construct) have been introduced to alleviate these problems, reality remains that many documents do not make use of metadata tags or, even worse, make use of them incorrectly and provide misleading information.

By not having metadata provided for us, or by deciding not to trust it, we are left with deducing information from the text itself. This is the domain of natural language processing (NLP) and text mining. This article deals with one aspect of text mining, namely telling which language (or languages) is a given Web page written in.

2 Related work

A general paradigm in automated language identification is to create language models during a training phase and compare input document against these mod-

Business Specific Online Information Extraction from German Websites

Yeong Su Lee and Michaela Geierhos

CIS, University of Munich, Germany

Abstract. This paper presents a system that uses the domain name of a German business website to locate its information pages (e.g. company profile, contact page, imprint) and then identifies business specific information. We therefore concentrate on the extraction of characteristic vocabulary like company names, addresses, contact details, CEOs, etc. Above all, we interpret the HTML structure of documents and analyze some contextual facts to transform the unstructured web pages into structured forms. Our approach is quite robust in variability of the DOM, upgradeable and keeps data up-to-date. The evaluation experiments show high efficiency of information access to the generated data. Hence, the developed technique is adaptive to non-German websites with slight language-specific modifications, and experimental results on real-life websites confirm the feasibility of the approach.

1 Introduction

With the expansion of the Web, the demand for targeted information extraction is continuously growing. There are many services on the Web providing industry sector information or performing job search tasks. For these purposes, the data used must be first manually collected and therefore features several sources of error, e.g. spelling mistakes, incomplete database entries, etc. Moreover, this process is extremely time-consuming and updating the data then requires a rollback of the full process. Automating these tasks will help to extract the business specific information quickly and maintain the data up-to-date.

The standard approach of business-related information retrieval disregards the relationship between the domain name and organization-specific content of a website, but concentrates on the structural aspect of company information [1]. Only a few studies restrict the information extraction task to certain domain names [2–4]. They extract company profiles by limiting their research on locating products and other features while analyzing the format of HTML tables for structured data and trying to find the phrase patterns for unstructured texts [2]. Others examine the presentation ontology for extracting organization-specific data such as contact details and product information concentrating on the differences in the presentation manner of formatted company profiles versus plain text profiles [3]. But company information extraction can also be extended to different resources and incorporates meta tags as well as plain texts and structured data [4].

Low-Cost Supervision for Multiple-Source Attribute Extraction

Joseph Reisinger[1]* and Marius Paşca[2]

[1] University of Texas at Austin, Austin, Texas, joeraii@cs.utexas.edu
[2] Google Inc., Mountain View, California, mars@google.com

Abstract. Previous studies on extracting class attributes from unstructured text consider either Web documents or query logs as the source of textual data. Web search queries have been shown to yield attributes of higher quality. However, since many relevant attributes found in Web documents occur infrequently in query logs, Web documents remain an important source for extraction. In this paper, we introduce Bootstrapped Web Search (BWS) extraction, the first approach to extracting class attributes simultaneously from both sources. Extraction is guided by a small set of seed attributes and does not rely on further domain-specific knowledge. BWS is shown to improve extraction precision and also to improve attribute relevance across 40 test classes.

1 Introduction

Class attributes capture quantifiable properties (e.g., *hiking trails*, *entrance fee* and *elevation*), of given classes of instances (e.g., *NationalPark*), and thus potentially serve as a skeleton towards constructing large-scale knowledge bases automatically. Previous work on extracting class attributes from unstructured text consider either Web documents [1] or query logs [2] as the extraction source. In this paper, we develop *Bootstrapped Web Search* (BWS), a method for combining Web documents and query logs as textual data sources that may contain class attributes. Web documents have textual content of higher semantic quality, convey information directly in natural language rather than through sets of keywords, and contain more raw textual data. In contrast, search queries are usually ambiguous, short, keyword-based approximations of often-underspecified user information needs. Previous work has shown, however, that extraction from query logs yields significantly higher precision than extraction from Web documents [2].

BWS is a generic method for multiple-source class attribute extraction that allows for corpora with varying levels of extraction precision to be combined favorably. It requires no supervision other than a small set of seed attributes for each semantic class. We test this method by combining query log and Web document corpora, leveraging their unique strengths in order to improve coverage and precision. Using BWS, extracted attributes from classes pertaining to various domains of interest to Web search users yield accuracy exceeding current state of the art using either Web documents or search queries alone.

* Contributions made during an internship at Google.

An Integrated Architecture for Processing Business Documents in Turkish

Serif Adali [1], A. Coskun Sonmez [2], Mehmet Gokturk [3]

[1] Istanbul Technical University, Department of Computer Engineering,
34469, Istanbul, Turkey, serifadali@yahoo.com
[2] Yildiz Technical University, Department of Computer Engineering,
34349, Istanbul, Turkey, acsonmez@ce.yildiz.edu.tr
[3] Gebze Institute of Technology, Department of Computer Engineering,
41400, Kocaeli, Turkey, gokturk@gyte.edu.tr

Abstract. This paper covers the first research activity in the field of automatic processing of business documents in Turkish. In contrast to traditional information extraction systems which process input text as a linear sequence of words and focus on semantic aspects, proposed approach doesn't ignore document layout information and benefits hints provided by layout analysis. In addition, approach not only checks relations of entities across document for verifying its integrity, but also verifies extracted information against real word data (e.g. customer database). This rule-based approach uses a morphological analyzer for Turkish, a lexicon integrated domain ontology, a document layout analyzer, an extraction ontology and a template mining module. Based on extraction ontology, conceptual sentence analysis increases portability which requires only domain concepts when compared to information extraction systems that rely on large set of linguistic patterns.

1 Introduction

Even though there has been an on-going effort for eliminating free-formatted text documents for almost thirty years, certain forms of communication continue to be unstructured such as fax and e-mail, where free-formatted text needs to be interpreted in order to understand its meaning and extract necessary information. Proposed integrated architecture has been tested for processing business documents in Turkish. The test data has been created based on sample documents provided by a local bank, by generating fictitious customer data (e.g. names, account number). Test documents are assumed to be optically recognized (OCR) and free of spell checking errors. Input document is treated as a combination of document model and event concept, where entities within document are cross-related to each other, and document can be correctly understood if certain clues are provided by the document layout analysis. Proposed integrated architecture combines document layout analysis, ontology-based information extraction and morphological analyzer for Turkish. Information Extraction (IE) is generally defined as a form of natural language processing, in

Detecting Protein-Protein Interactions in Biomedical Texts using a Parser and Linguistic Resources

Gerold Schneider, Kaarel Kaljurand, and Fabio Rinaldi

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

Abstract. We describe the task of automatically detecting interactions between proteins in biomedical literature. We use a syntactic parser, a corpus annotated for proteins, and manual decisions as training material. After automatically parsing the GENIA corpus, which is manually annotated for proteins, all syntactic paths between proteins are extracted. These syntactic paths are manually disambiguated between meaningful paths and irrelevant paths. Meaningful paths are paths that express an interaction between the syntactically connected proteins, irrelevant paths are paths that do not convey any interaction.

The resource created by these manual decisions is used in two ways. First, words that appear frequently inside a meaningful paths are learnt using simple machine learning. Second, these resources are applied to the task of automatically detecting interactions between proteins in biomedical literature. We use the IntAct corpus as an application corpus.

After detecting proteins in the IntAct texts, we automatically parse them and classify the syntactic paths between them using the meaningful paths from the resource created on GENIA and addressing sparse data problems by shortening the paths based on the words frequently appearing inside the meaningful paths, so-called transparent words.

We conduct an evaluation showing that we achieve acceptable recall and good precision, and we discuss the importance of transparent words for the task.

1 Introduction

Scientific articles reporting results of biomedical studies are growing exponentially in number[1]. Publically available literature services such as Pubmed (http://pubmed.gov) already contain more than 17 million articles. Even for the expert it has become difficult to keep an overview of new results. Fully or partly automated systems that extract biological knowledge from text have thus become a necessity. Particularly, knowledge about protein-protein interactions

[1] This research is partially funded by the Swiss National Science Foundation (grant 100014-118396/1). Additional support is provided by Novartis Pharma AG, NITAS, Text Mining Services, CH-4002, Basel, Switzerland.

Learning to learn biological relations from a small training set

Laura Alonso i Alemany, Santiago Bruno

NLP Group
Facultad de Matemática Astronomía y Física (FaMAF), UNC
Córdoba, Argentina

Abstract. In this paper we present different ways to improve a basic machine learning approach to identify relations between biological named entities as annotated in the Genia corpus.
The main difficulty with learning from the Genia event-annotated corpus is the small amount of examples that are available for each relation type. We compare different ways to address the data sparseness problem: using the corpus as the initial seed of a bootstrapping procedure, generalizing classes of relations via the Genia ontology and generalizing classes via clustering.

1 Introduction and Motivation

The huge amount of biomedical research papers available nowadays makes intelligent information access a necessity. In this paper we develop a method to assist information retrieval for highly focused information needs. In particular, we develop a module to detect relations between biological entities in free text. The work presented here is part of the MicroBio project[1], which aims to build a system for information retrieval of biomedical research papers. This module will be inserted in a wider system for information retrieval of biomedical research papers. This module will provide the capability of querying a database to find documents containing a particular relation between biological entities.

Given the importance of biomedical research, much effort has been devoted to the problem of biomedical information access. Very good results have been obtained for the recognition and classification of biological named entities (Bio-NER). In contrast, approaches to the problem of discovering relations between biological entities have been more rare and less successful, probably because the problem is more complex.

To begin with, there are less and smaller corpora annotated with relations, which constitutes a considerable bottleneck for supervised machine learning approaches to the problem. Unsupervised machine learning approaches do not suffer from lack of training data, but are normally incapable of detecting fine-grained distinctions in relations. As for symbolic approaches, they suffer from

[1] http://www.microbioamsud.net

Using a Bigram Event Model to Predict Causal Potential

Brandon Beamer and Roxana Girju

University of Illinois at Urbana-Champaign
{bbeamer,girju}@illinois.edu

Abstract. This paper addresses the problem of causal knowledge discovery. Using online screenplays, we generate a corpus of temporally ordered events. We then introduce a measure we call *causal potential* which is easily calculated with statistics gathered over the corpus and show that this measure is highly correlated with an event pair's tendency of encoding a causal relation. We suggest that causal potential can be used in systems whose task is to determine the existence of causality between temporally adjacent events, when critical context is either missing or unreliable. Moreover, we argue that our model should therefore be used as a baseline for standard supervised models which take into account contextual information.

1 Introduction

Automatic recognition and extraction of causal event sequences from text is a crucial task for many Computational Linguistics applications. It is a prerequisite in text coherence, entailment, question answering, and information retrieval (Goldman et al., 1999; Khoo et al., 2001; Girju, 2003). Put simply, it is a prerequisite to perform textual *reasoning*.

Whether two textual units (words, phrases or sentences) are in a causal relationship is largely dependent on context. But that is not to say that such pairs in general have no statistical tendencies when it comes to causality.

This paper describes a knowledge-poor unsupervised model relying on a statistical measure we call *causal potential*. Our focus is on event sequences as expressed by consecutive verbs in a discourse (event pairs). Event pairs with a high causal potential can be interpreted as being more likely to occur in causal contexts than events with low causal potential. This measure can then be used in more complex systems to gain causal intuitions in situations when context is scarce or unreliable. Therefore, we argue that our model should be used as a baseline by standard supervised models which take into account contextual information.

In this paper we evaluate our measure of causal potential and show that it correlates highly with human observances of causal events in text.

Semantic-Based Temporal Text-Rule Mining

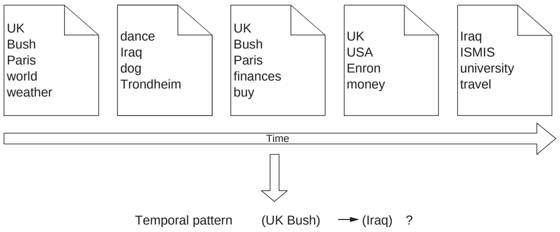
Kjetil Nørvåg* and Ole Kristian Fivelstad

Dept. of Computer Science, Norwegian University of Science and Technology
Trondheim, Norway

Abstract. In many contexts today, documents are available in a number of versions. In addition to *explicit knowledge* that can be queried/searched in documents, these documents also contain *implicit knowledge* that can be found by text mining. In this paper we will study association rule mining of temporal document collections, and extend previous work within the area by 1) performing mining based on *semantics* as well as 2) studying the impact of appropriate techniques for ranking of rules.

1 Introduction

In many contexts today, documents are available in a number of versions. Examples include web newspapers and health records, where a number of timestamped document versions are available. In addition to *explicit knowledge* that can be queried/searched in documents, these documents also contain *implicit knowledge*. One category is inter-document knowledge that can be found by conventional text-mining techniques. However, with many versions available there is also the possibility of finding *inter-version knowledge*. An example of an application is given in the figure below, where a number of document versions are available, and where the aim is to find and/or verify temporal patterns:



In the example above, one possible temporal rule is the terms[1] *UK* and *Bush* appearing in one version means a high probability of *Iraq* to appear in one of the following versions.

How to mine association rules in temporal document collection has been previously described in [16]. In the previous work, the rule mining was performed on *words* extracted from the documents, and ranking of rules (in order to find the most interesting

[*] E-mail of contact author: Kjetil.Norvag@idi.ntnu.no

[1] A term can be a single word as well as multiword phrase.

Generating Executable Scenarios from Natural Language

Michal Gordon and David Harel

The Weizmann Institute of Science, Rehovot, 76100, Israel
{michal.gordon,dharel}@weizmann.ac.il

Abstract. Bridging the gap between the specification of software requirements and actual execution of the behavior of the specified system has been the target of much research in recent years. We have created a natural language interface, which, for a useful class of systems, yields the automatic production of executable code from structured requirements. In this paper we describe how our method uses static and dynamic grammar for generating live sequence charts (LSCs), that constitute a powerful executable extension of sequence diagrams for reactive systems. We have implemented an automatic translation from controlled natural language requirements into LSCs, and we demonstrate it on two sample reactive systems.

1 Introduction

Live Sequence Charts are a visual formalism that describes natural "pieces" of behavior and are similar to telling someone what they may and may not do, and under what conditions. The question we want to address here is this: can we capture the requirements for a dynamic system in a far more natural style than is common? We want a style that is intuitive and less formal, and which can also serve as the system's executable behavioral description [1].

To be able to specify behavior in a natural style, one would require a simple way to specify pieces of requirements for complex behavior, without having to explicitly, and manually, integrate the requirements into a coherent design. In [2], the mechanism of *play-in* was suggested as a means for making programming practical for lay-people. In this approach, the user specifies scenarios by playing them in directly from a graphical user interface (GUI) of the system being developed. The developer interacts with the GUI that represents the objects in the system, still a behavior-less system, in order to show, or teach, the scenario-based behavior of the system by example (e.g., by clicking buttons, changing properties or sending messages). As a result, the system generates automatically, and on the fly, live sequence charts (LSCs) [3], a variant of UML sequence diagrams [4] that capture the behavior and interaction between the environment and the system or between the system's parts. In the current work we present an initial natural language interface that generates LSCs from structured English requirements.

Determining the Polarity and Source of Opinions Expressed in Political Debates

Alexandra Balahur, Zornitsa Kozareva, Andrés Montoyo

University of Alicante, Departament of Software and Computing Systems,
Apartado de Correos 99, E-03080 Alicante, Spain
{abalahur, zkozareva, montoyo}@dlsi.ua.es

Abstract. In this paper we investigate different approaches we developed in order to classify opinion and discover opinion sources from text, using affect, opinion and attitude lexicon. We apply these approaches on the discussion topics contained in a corpus of American Congressional speech data. We propose three approaches to classifying opinion at the speech segment level, firstly using similarity measures to the affect, opinion and attitude lexicon, secondly dependency analysis and thirdly SVM machine learning. Further, we study the impact of taking into consideration the source of opinion and the consistency in the opinion expressed, and propose three methods to classify opinion at the speaker intervention level, showing improvements over the classification of individual text segments. Finally, we propose a method to identify the party the opinion belongs to, through the identification of specific affective and non-affective lexicon used in the argumentations. We present the results obtained when evaluating the different methods we developed, together with a discussion on the issues encountered and some possible solutions. We conclude that, even at a more general level, our approach performs better than trained classifiers on specific data.

Keywords: opinion mining, opinion source mining, LSA, political discourse.

1 Introduction

Most people, at the time of taking a decision, base their choice on a series of arguments, which are rational and/or emotional. For example, the factors influencing the purchase of a certain digital camera over another might take into consideration rational arguments, such as price, performance, but also emotional arguments, such as brand reputation and opinion expressed by other buyers of that camera. On the other hand, the decision to vote for a political party or a given candidate, while mostly based on the political beliefs one has, can be strongly influenced by whether or not, from the actions performed by the party or candidate, those political ideas are respected.

Recent years have marked the beginning and expansion of the social web, in which people freely express and respond to opinion on a whole variety of topics. Moreover, at the time of taking a decision, more and more people search for information and

Query Translation and Expansion for Searching Normal and OCR-Degraded Arabic Text

Tarek Elghazaly and Aly Fahmy

Faculty of Computers and Information, Cairo University, Giza, Egypt,
{t.elghazaly, a.fahmy}@fci-cu.edu.eg

Abstract. This paper provides a novel model for English/Arabic Query Translation to search Arabic text, and then expands the Arabic query to handle Arabic OCR-Degraded Text. This includes detection and translation of word collocations, translating single words, transliterating names, and disambiguating translation and transliteration through different approaches. It also expands the query with the expected OCR-Errors that are generated from the Arabic OCR-Errors simulation model which proposed inside the paper. The query translation and expansion model has been supported by different libraries proposed in the paper like a Word Collocations Dictionary, Single Words Dictionaries, a Modern Arabic corpus, and other tools. The model gives high accuracy in translating the Queries from English to Arabic solving the translation and transliteration ambiguities and with orthographic query expansion; it gives high degree of accuracy in handling OCR errors.

Keywords: Query Translation, Orthographic Query Expansion, Cross Language Information Retrieval, Arabic OCR-Degraded Text, Arabic Corpus.

1 Introduction

The importance of Cross Language Information Retrieval (CLIR) appears clearly when we consider a case like the Library of Congress [1] which has more than 134 million items and approximately half of the library's book and serial collections are in 460 languages other than English. When people like to retrieve the whole set of documents that represent some interest, they have to repeat search process in each language. Furthermore, as a big number of books and documents are available only in print especially the Arabic ones, they are not 'full text' searchable and they need applying the Arabic OCR process whose accuracy is far from perfect [2]. The goal of this paper is to provide a solid English/Arabic query translation and expansion model to search both normal and OCR-Degraded Arabic Text.

The outline of this paper is as follows: The previous work is reviewed in Section 2. The proposed work is presented in the next sections. Arabic words formalization, normalization and stemming are presented in Section 3. Corpus and Dictionaries are presented in Section 4 and 5. In Section 6 & 7 the work done for CLIR through Query Translation and expansion respectively is detailed, followed by the experimental results and the conclusions in Sections 8 & 9.

NLP for Shallow Question Answering of Legal Documents Using Graphs*

Alfredo Monroy[1], Hiram Calvo[1,2], Alexander Gelbukh[1]

[1]Center for Computing Research, National Polytechnic Institute
Mexico City, 07738, Mexico
alopezm301@ipn.mx; hcalvo@cic.ipn.mx; gelbukh@gelbukh.com

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

Abstract. Previous work has shown that modeling relationships between articles of a regulation as vertices of a graph network works twice as better than traditional information retrieval systems for returning articles relevant to the question. In this work we experiment by using natural language techniques such as lemmatizing and using manual and automatic thesauri for improving question based document retrieval. For the construction of the graph, we follow the approach of representing the set of all the articles as a graph; the question is split in two parts, and each of them is added as part of the graph. Then several paths are constructed from part A of the question to part B, so that the shortest path contains the relevant articles to the question. We evaluate our method comparing the answers given by a traditional information retrieval system—vector space model adjusted for article retrieval, instead of document retrieval—and the answers to 21 questions given manually by the general lawyer of the National Polytechnic Institute, based on 25 different regulations (academy regulation, scholarships regulation, postgraduate studies regulation, etc.); with the answer of our system based on the same set of regulations. We found that lemmatizing increases performance in around 10%, while the use of thesaurus has a low impact.

1 Introduction

Previous work [20] has shown that modelling relationships between articles of a regulation as vertices of a graph network works twice as better than traditional information retrieval systems for returning articles relevant to the question. Despite being that approach language independent, in this work we experiment by using natural language techniques such as lemmatizing and using manual and automatic thesauri for improving question based document retrieval. We focus in Spanish language. For automatic thesaurus we used a distributional thesaurus [19], and for the manual thesaurus we used a human oriented dictionary (Anaya) [21]. The advantage of using a distributional thesaurus is that the approach remains language independent—not being the

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

Semantic Clustering for a Functional Text Classification Task

Thomas Lippincott and Rebecca Passonneau

Columbia University
Department of Computer Science
Center for Computational Learning Systems
New York, NY USA
tom,becky@cs.columbia.edu

Abstract. We describe a semantic clustering method designed to address shortcomings in the common bag-of-words document representation for functional semantic classification tasks. The method uses WordNet-based distance metrics to construct a similarity matrix, and expectation maximization to find and represent clusters of semantically-related terms. Using these clusters as features for machine learning helps maintain performance across distinct, domain-specific vocabularies while reducing the size of the document representation. We present promising results along these lines, and evaluate several algorithms and parameters that influence machine learning performance. We discuss limitations of the study and future work for optimizing and evaluating the method.

1 Introduction

The bag-of-words document representation achieves excellent performance on many machine learning tasks. However, this performance can be sensitive to the changes in vocabulary that occur when the training data cannot be reasonably expected to be representative of all the potential testing data. In this situation, it may be possible to exploit higher-level relationships between the vocabularies by consolidating and generalizing the specific bag-of-words features into a smaller number of semantic clusters using an external semantic resource. This would have the dual benefits of retaining performance across domains and reducing the dimensionality of the document representation.

The approach presented here was developed in response to characteristics of the machine learning phase of the ANTArT(1), a component of the CLiMB research project (2). CLiMB developed a toolkit for image catalogers that facilitates harvesting descriptive meta-data from scholarly text for annotating digital image collections. ANTArT, collaborating with experts in art history and image cataloging, developed a set of functional semantic labels to characterize how art-historical texts function with respect to their associated images (see Table 4 for the complete list), drawing on texts from four art-historical periods.

Three data sets were prepared from art history texts covering two time periods: Near Eastern art (two data sets), and Medieval art (one data set). Each

Reducing the Plagiarism Detection Search Space on the basis of the Kullback-Leibler Distance

Alberto Barrón-Cedeño, Paolo Rosso, and José-Miguel Benedí

Department of Information Systems and Computation,
Universidad Politécnica de Valencia,
Valencia 46022, Camino de Vera s/n, Spain
{lbarron, prosso, jbenedi}@dsic.upv.es
http://www.dsic.upv.es/grupos/nle/

Abstract. Automatic plagiarism detection considering a reference corpus compares a suspicious text to a set of original documents in order to relate the plagiarised fragments to their potential source. Publications on this task often assume that the search space (the set of reference documents) is a narrow set where any search strategy will produce a good output in a short time. However, this is not always true. Reference corpora are often composed of a big set of original documents where a simple exhaustive search strategy becomes practically impossible.
Before carrying out an exhaustive search, it is necessary to reduce the search space, represented by the documents in the reference corpus, as much as possible. Our experiments with the METER corpus show that a previous search space reduction stage, based on the Kullback-Leibler symmetric distance, reduces the search process time dramatically. Additionally, it improves the Precision and Recall obtained by a search strategy based on the exhaustive comparison of word n-grams.

1 Introduction

The easy access to a wide range of information via electronic resources such as the Web, has favoured the increase of plagiarism cases. When talking about text, plagiarism means to use text written by other people (even adapting it by rewording, insertion or deletion), without any credit or citation. In fact, the reuse of self-written text is often considered as self-plagiarism.

Plagiarism detection with reference tries to find the source of the potentially plagiarised fragments from a suspicious document in a set of reference documents. Some techniques based on the exhaustive comparison of suspicious and original documents have already been developed. These techniques apply comparison of sentences [7], structure of documents [15] and entire documents [10]. Examples of the used comparison strategies are dot plot [15] and n-grams [10].

One of the main difficulties in this task is the great size of the search space, i.e., the reference documents. To our knowledge, this problem has not been studied deeply enough, neither there are published papers on this issue. Given a suspicious document, our current research is precisely oriented to the reduction

Empirical Paraphrasing of Modern Greek Text in Two Phases: An Application to Steganography

Katia Lida Kermanidis and Emmanouil Magkos

Ionian University, Department of Informatics
7 Pl. Tsirigoti, 49100, Corfu, Greece
{kerman, emagos}@ionio.gr

Abstract. This paper describes the application of paraphrasing to steganography, using Modern Greek text as the cover medium. Paraphrases are learned in two phases: a set of shallow empirical rules are applied to every input sentence, leading to an initial pool of paraphrases. The pool is then filtered through supervised learning techniques. The syntactic transformations are shallow and require minimal linguistic resources, allowing the methodology to be easily portable to other inflectional languages. A secret key shared between two communicating parties helps them agree on one chosen paraphrase, the presence of which (or not) represents a binary bit of hidden information. The ability to simultaneously apply more than one rules, and each rule more than one times, to an input sentence increases the paraphrase pool size, ensuring thereby steganographic security.

Keywords: paraphrasing, shallow parsing, supervised learning, steganography

1 Introduction

Given an original sentence, that conveys a specific meaning, paraphrasing means expressing the same meaning using a different set of words or a different syntactic structure. Significant research effort has been put into the identification as well as the generation of paraphrases. Paraphrasing has been used extensively for educational purposes in language learning, as well as in several NLP tasks like text summarization [4], question answering [6] and natural language generation. Recently it has found yet another use in steganography.

Regarding paraphrase identification, previous approaches have utilized supervised [8] or unsupervised ([2][3]) machine learning techniques. The authors in [13] use named entity recognition in newswire articles from different newspapers to detect pairs of sentences that discuss the same topic and find sentence parts that are paraphrases. Regarding paraphrase generation, the use of finite state automata [10], i.e. paraphrase lattices, has been proposed, as well as the application of empirical rules [9] and statistical machine translation techniques [12].

In the present work, paraphrases of Modern Greek free text are learned in two phases. Henceforth, the term ‘paraphrasing’ will stand for shallow syntactic transformations, i.e. swaps of consecutive phrasal chunks. Modern Greek is suitable

BorderFlow: A Local Graph Clustering Algorithm for Natural Language Processing

Axel-Cyrille Ngonga Ngomo[1] and Frank Schumacher[1]

Department of Business Information Systems
University of Leipzig
Johannisgasse 26, Leipzig D-04103, Germany
{ngonga|schumacher}@informatik.uni-leipzig.de,
WWW home page: http://bis.uni-leipzig.de/

Abstract. In this paper, we introduce BorderFlow, a novel local graph clustering algorithm, and its application to natural language processing problems. For this purpose, we first present a formal description of the algorithm. Then, we use BorderFlow to cluster large graphs and to extract concepts from word similarity graphs. The clustering of large graphs is carried out on graphs extracted from the Wikipedia Category Graph. The subsequent low-bias extraction of concepts is carried out on two data sets consisting of noisy and clean data. We show that BorderFlow efficiently computes clusters of high quality and purity. Therefore, BorderFlow can be integrated in several other natural language processing applications.

1 Introduction

Graph-theoretical models and algorithms have been successfully applied to natural language processing (NLP) tasks over the past years. Especially, graph clustering has been applied to areas as different as language separation [3], lexical acquisition [9] and word sense disambiguation [12]. The graphs generated in NLP are usually large. Therefore, most global graph clustering approaches fail when applied to NLP problems. Furthermore, certain applications (such as concept extraction) require algorithms able to generate a soft clustering. In this paper, we present a novel local graph clustering algorithm called BorderFlow, which is designed especially to compute a soft clustering of large graphs. We apply BorderFlow to two NLP-relevant tasks, i.e., clustering large graphs and concept extraction. We show that our algorithm can be effectively used to tackle these two tasks by providing quantitative and qualitative evaluations of our results.

This paper is structured as follows: in the next section, we describe BorderFlow formally. Thereafter, we present our experiments and results. First, we present the results obtained using BorderFlow on three large similarity graphs extracted from the Wikipedia Category Graph (WCG). By these means, we show that BorderFlow can efficiently handle large graphs. Second, we use BorderFlow to extract domain-specific concepts from two different corpora and show that it computes concepts of high purity. Subsequently, we conclude by discussing possible extensions and applications of BorderFlow.

Generalized Mongue-Elkan Method for Approximate Text String Comparison

Sergio Jimenez,[1] Claudia Becerra,[1] Alexander Gelbukh,[2] and Fabio Gonzalez[1]

[1]Intelligent Systems Laboratory (LISI)
Systems and Industrial Engineering Department
National University of Colombia
[2]Natural Language Laboratory
Center for Computing Research (CIC)
National Polytechnic Institute (IPN), Mexico

Abstract. The Mongue-Elkan method is a general text string comparison method based on an internal character-based similarity measure (e.g. edit distance) combined with a token level (*i.e.* word level) similarity measure. We propose a generalization of this method based on the notion of the generalized arithmetic mean instead of the simple average used in the expression to calculate the Monge-Elkan method. The experiments carried out with 12 well-known name-matching data sets show that the proposed approach outperforms the original Monge-Elkan method when character-based measures are used to compare tokens.

1 Introduction

Approximate string similarity measures are used in many natural language processing (NLP) tasks such as term identification in information extraction, information retrieval, word sense disambiguation, etc. Probably the most well known character-level measure is the *edit distance* proposed by Levenshtein in 1965 [11]. The character-based measures consider the strings to be compared merely as character sequences, which makes this approach affordable when the strings to be compared are single words having misspellings, typographical errors, OCR errors, or even some morphological variations. However, in most human languages, character sequences are split into words (tokens). This property of natural language texts is exploited by token-based measures such as the resemblance coefficients [1] (e.g., *Jaccard*, *Dice*, overlap). The token-based measures compare text strings as sequences of tokens instead of sequences of characters. Such an approach is successful when it is used to compare text strings with many tokens and with different order of the tokens or missing tokens.

All previously mentioned measures based on character or tokens are static. Static string-similarity metrics as they were defined by Bilenko *et al.* [3] are those that compare two character or token sequences in an algorithmic way using solely the information contained into the sequences. Most of the character-based measures are static, that is, the characters into two strings are compared among them with a strategy in order to return a final similarity value. Some

Estimating Risk of Picking a Sentence for Document Summarization

Chandan Kumar, Prasad Pingali, and Vasudeva Varma

Language Technologies Research Centre,
International Institute of Information Technology,
Hyderabad, India
chandan_kumar@research.iiit.ac.in
{pvvpr,vv}@iiit.ac.in

Abstract. Automatic Document summarization is proving to be an increasingly important task to overcome the information overload. The primary task of document summarization process is to pick subset of sentences as a representative of whole document set. We treat this as a decision making problem and estimate the risk involve in making this decision. We calculate the risk of information loss associated with each sentence and extract sentences based on ascending order of their risk. The experimental result shows that the proposed approach performs better than various state of the art approaches.

1 Introduction

Automatic document summarization is extremely helpful in saving time and efforts of the users by helping in tackling the information overload problems. The focus in automatic summarization has been shifted from single document summarization to more complex and challenging problem of multi-document summarization. The goal here is to produce a single text as a compressed version of a given set of documents related to a particular topic with all and only the relevant information.

There are two kinds of approaches to document summarization: abstraction and extraction. Even though efforts have been put to generate an abstract summary that requires using heavy machinery from natural language processing, including grammars and lexicons for parsing and generation, extraction is still the most feasible approach, and most of recent works in this area are based on extraction. Extraction is the process of selecting important units from the original document and concatenating them into a shorter form as summary. Extractive approach to summarization can employ various levels of granularity, e.g., keyword, sentence, or paragraph. Most research concentrates on sentence-extraction because the readability of a list of keywords is typically low while paragraphs are unlikely to cover the information content of a document given summary space constraints.

In this paper, we address the problem of generic multi-document summarization through a sentence extractive procedure. Here the task is to pick subset of

The Decomposition of Human-Written Book Summaries

Hakan Ceylan and Rada Mihalcea

University of North Texas
Computer Science Department
{hakan,rada}@unt.edu

Abstract. *In this paper, we evaluate the extent to which human-written book summaries can be obtained through cut-and-paste operations from the original book. We analyze the effect of the parameters involved in the decomposition algorithm, and highlight the distinctions in coverage obtained for different summary types.*

1 Introduction

Books represent one of the oldest forms of written communication. Despite this fact, given that a large fraction of the electronic documents available online and elsewhere consist of short texts such as Web pages or news articles, the focus of natural language processing techniques to date has been on the automation of methods targeting short documents. We are witnessing however a change: an increasingly larger number of books become available in electronic format, in projects such as Gutenberg, Google Books, or the Million Books project. Similarly, a large number of the books published in recent years are often available in electronic format. Thus, the need for language processing techniques able to handle very large documents such as books is becoming increasingly important.

In this paper, we focus on the problem of book summarization. In particular, we address the first step in automatic summarization, and analyze the extent to which human-written summaries of books can be obtained through extractive methods. We use a decomposition algorithm to automatically identify matches between sentences in the summary and sentences in the book, and thus determine the potential coverage of extractive summarization.

Our work is inspired by the decomposition algorithm previously proposed by Jing & McKeown for single-document summarization [4]. In this paper, we study the applicability of the algorithm to book summaries, and analyze the effect of its various parameters on the coverage of the decomposition. We show that even for long documents such as books, a significant number of summary sentences can be obtained through cut-and-paste operations from the original book. In turn, this coverage depends on the type of summaries being analyzed, with significant differences observed between objective (plot) summaries and interpretative summaries.

Linguistic Ethnography: Identifying Dominant Word Classes in Text

Rada Mihalcea[1,2], Stephen Pulman[2]

[1] Computer Science Department, University of North Texas
rada@cs.unt.edu
[2] Computational Linguistics Group, Oxford University
sgp@clg.ox.ac.uk

Abstract. In this paper, we propose a method for ”linguistic ethnography” – a general mechanism for characterising texts with respect to the dominance of certain classes of words. Using humour as a case study, we explore the automatic learning of salient word classes, including semantic classes (e.g., person, animal), psycholinguistic classes (e.g., tentative, cause), and affective load (e.g., anger, happiness). We measure the reliability of the derived word classes and their associated dominance scores by showing significant correlation across different corpora.

1 Introduction

Text classification is an area in natural language processing that has received a significant amount of interest from both the research and industrial communities, with numerous applications ranging from spam detection and Web directory categorization [4], to sentiment and subjectivity classification [17], emotion recognition [14], gender identification [3] or humour recognition [6]. The task is defined as the automatic identification and labeling of texts that share certain properties, be that a common topic (e.g., “arts”), a common author (e.g., female-authored texts), or a certain feature of the text (e.g., humorous texts).

While there are a number of text classification algorithms that have been proposed to date, there are only a handful of techniques that have been developed to identify the characteristics that are shared by the texts in a given class. Most of the work in this area has focused on the use of weights associated with the words in the text, by using metrics such as tf.idf or information gain, but no attempts have been made to systematically identify broader patterns or word classes that are common in these texts. The relatively small amount of work in this area is understandable since, from a practical perspective, the accurate classification of texts is more important than the identification of general word classes that are specific to the texts in one category.

When the goal however is to *understand the characteristics* of a certain type of text, in order to gain a better understanding of the properties or behaviours modeled by those texts (such as happiness, humour, or gender), then the systematic identification of broad word classes characteristic to the given type of text is considerably more insightful than a mere figure reflecting the accuracy of a text classifier.

Given a collection of texts, characterised by a certain property that is shared by all the texts in the collection, we introduce a method to automatically discover the classes