LANGUAGE FORUM
VOL. 34, NO. 1, JAN-JUN 2008

Data Model for a Lexical Resource Based on Lexical Functions

SOCORRO BERNARDOS
Universidad Politécnica de Madrid, Spain

MARÍA A. BARRIOS
Universidad Complutense de Madrid, Spain

ABSTRACT

The purpose of this paper is to describe a data model that has been used in the construction of a lexical database for Spanish. This lexical resource takes the Meaning-Text Theory (MTT) as the theoretical basis, especially lexical functions (LF), which associate a lexical unit (LU) to values that express a specific meaning related to that LF for that LU. We have developed a database, called BaDELE3000, for the 3,000 most frequently used nouns in Iberian Spanish. We followed a systematic process for the design, so that the lexical data are well structured and separated from the applications that might use them. This way, the features of the data model and the subsequent database make them useful for different purposes such as word sense disambiguation, machine translation and text generation.

1. INTRODUCTION

In 2002 the Laboratory of Computational Linguistics of Moscow (Apresjan et al. 2002), (Apresjan et al. 2003) *CALLEX*, a computer-assisted language learning tool for Russian, English and German which helps the learners of one of this languages to extend their lexical knowledge by means of five types of games. This tool is based on the general linguistic framework of the Meaning-Text Theory (MTT), proposed by Igor Mel'čuk (2003), especially on Lexical Functions (LFs), which are used to represent lexical relations: paradigmatic (synonyms, antonyms, hyperonyms, etc.) and syntagmatic (collocations).

The adaptation of that tool for Spanish, called *CALLEX-ESP*, is under development. This paper deals with the data model created for

LANGUAGE FORUM
VOL. 34, NO. 1, JAN-JUN 2008

Ontology-Supported Automated Mark Up of Affective Information in Texts

VIRGINIA FRANCISCO
PABLO GERVÁS
Universidad Complutense de Madrid, Spain

ABSTRACT

This paper presents the application of an ontology of emotions to an existing approach for the automated mark up of affective information in texts. The emotional ontology has three main applications in this system: to select the most specific emotion which represents the affective information of the sentences from the probability that each word in the sentence has of indicating different emotions, to establish what emotion should be assigned to a sentence given the set of emotional assignments suggested by a group of evaluators, and to determine if the emotion assigned to the sentence during automated mark up is correct. The enhanced markup system has been tested and the results show improvements with respect to the previous version.

1. INTRODUCTION

Human languages have produced extremely powerful labels for emotional states, for example, English provides at least 107 emotion-denoting adjectives and German at least 235. To consider each of these categories as individual labels, not related with any other category, produces an uncontrolled proliferation of labels which multiplies the complexity of tasks which involve the use of these labels. If the emotional categories were not individual isolated units but units related with each other this might simplify tasks such as comparing two different emotional labels or deciding which is the emotion that better represents the generalization of two different emotions. A taxonomy of emotional categories where emotional labels are structured and organized in levels, from the most general to the most specific might provide a very useful tool in the treatment of emotional categories.

If we have two different emotional labels and we want to compare them, the different granularity of the labels could be an important aspect

LANGUAGE FORUM
VOL. 34, NO. 1, JAN-JUN 2008

The Knowledge of Good and Evil: Multilingual Ideology Classification with PARAFAC2 and Machine Learning

PETER CHEW
Hoffmantown Church, Albuquerque, NM, USA

PHILIP KEGELMEYER
BRETT BADER
Sandia National Laboratories, Albuquerque, NM, USA

AHMED ABDELALI
New Mexico State University, Las Cruces, NM, USA

ABSTRACT

We explore the problem of automated classification of multilingual documents by ideology; that is, whether machine learning techniques can be applied to draw conclusions about the types of belief an author holds based just on text he/she produces. For training data, we use documents in a variety of languages. Some are collected from the world wide web by a "spider", while others are hand-picked "ideological" documents, such as the writings of Lenin and Hitler. The documents are projected into a single cross-language semantic space defined by the application of PARAFAC2 to a multi-parallel aligned corpus. We find that standard machine learning techniques can be used to distinguish "ideological" documents from general web content with a high rate of accuracy (higher than has been reported for sentiment analysis problems), but the accuracy is slightly lower in distinguishing between ideologies. We conclude by discussing the factors that may contribute to our findings.

1. INTRODUCTION

It is almost a truism by now to say that the amount of information available today as on-line text is huge. As the volume of available information has grown, so has interest in methods for automated

LANGUAGE FORUM
VOL. 34, NO. 1, JAN-JUN 2008

Image Specific Language Model: Comparing Language Models from Two Independent Distributions from FlickR and the Web

GREGORY GREFENSTETTE
GUILLAUME PITEL
Cea List, Fontenay Aux Roses, France

ABSTRACT

Here we present a study comparing vocabulary from two different language models: the language used by people to describe their pictures on Flickr, and unrestricted language found on the Web. We examine these two language models through the lens of the semantically typed vocabulary in the General Inquirer lexicon. Developed in the 1960s for computerized content analysis of text, it provides a number of semantic categories for each word it contains. We compare the relative frequencies of usage of these categories, and usage of individual word within these categories, in Flickr tags versus the Web, using the weirdness *metric from research in special languages. The distribution of words used in each independent set, image annotations and Web text, is found to be different as expected, and we describe the differences. We show, for example, that noun Flickr tags are much more preferred over adjectives and verbs, more than the Web would lead us to expect. Names for animals, humans, and weddings also appear more frequently in Flickr tag than on the Web. Exploring these two language models helps us to better understand the vocabulary specific to images, which should ultimately be helpful in determining vocabulary for automated image annotation.*

1. INTRODUCTION

Adam Kilgarriff (2005) has shown that language distributions over any two large text collections are always different, even if you just divide one corpus in two random parts. It would seem all the more obvious that distribution of words used by people use when they describe photos is different from that used when they write ordinary text. And though the existence of differences seems intuitively obvious, the

LANGUAGE FORUM
VOL. 34, NO. 1, JAN-JUN 2008

Revealing Granularity of Domain Terminology with Inductive Method of Model Self-Organization

NATALIA PONOMAREVA
Universidad Politécnica de Valencia, Spain,

ABSTRACT

The aim of this work is to suggest a method of domain terms selection for different granularity levels. First, we give a definition of corpus-based term granularity and propose entropy-based and standard deviation-based weighting schemes for its evaluation. A chosen term weighting scheme is a decisive factor of granularity approximation quality. We declare a hypothesis of how to reveal boundaries between different granularity levels using our modified version of the Inductive Method of Model Self-Organization. Although the suggested method demonstrates stability in the framework of our hypothesis some additional study of its reliability must be accomplished and other more precise weighting schemes should be applied.

1. INTRODUCTION

Evaluation of ontology granularity level is still a difficult and poorly reflected problem in literature, although the importance of its solution is indisputable. A notion of granularity is used in a very intuitive way, neither its formal definition nor a mode of its measuring has been proposed up to now. Ontology granularity can be considered from different aspects: either at a lexical level, which refers to granularity of ontology concepts or at a conceptual level where expressiveness of ontology properties and relations is in a center of investigation.

In this paper, we consider a problem of ontology granularity at a lexical level (the lowest level of expessiveness according to Ontology Summit 2007[1]). Therefore, we aim at evaluating only granularity of ontology concepts without taking into account other ontology components. In other words, approximation of ontology by a list of its concepts is realized. Although it is a very rude approximation we argue that it is a first step of ontology granularity evaluation.

LANGUAGE FORUM
VOL. 34, NO. 1, JAN-JUN 2008

Using Meaning Aspects for Word Sense Disambiguation

RONALD WINNEMÖLLER
University of Hamburg, Hamburg

ABSTRACT

In this paper, we describe the successful application of our approach to text sense representation in the area of word sense disambiguation. After an introduction to our underlying theory and after briefly describing our basic data structures and operations, we present our experimental results on the Senseval-3 English Lexical Sample task - compared to the somewhat lesser impressive results we achieved by using traditional word vector based data.

We hope that our work will eventually enable unified ways of achieving better results for certain text processing tasks in general.

1. INTRODUCTION

Word Sense Disambiguation (WSD) can be regarded as vital stepping stone for advanced text processing applications such as machine translation, semantic web search, automatic summarization and many more. By the term Word Sense Disambiguation, we mean the selection of a particular sense out of a set of sense definitions, according to the "calculated" meaning of a particular word within a short text fragment taken from a larger corpus.

Up to date there is a great variety of approaches to basic WSD problems, most of which use lexicosyntactic or vector space based methods (usually augmented by techniques related to dictionaries or ontologies). The majority of these methods require manually prepared training data in order to produce results of acceptable quality but there are some non-supervised approaches as well.

Even though great efforts were invested in WSD research, results are in general far from being perfect (best precision nowadays is around 70% for supervised systems, non-supervised systems usually are much worse (Mihalcea et al. 2004).

LANGUAGE FORUM
VOL. 34, NO. 1, JAN-JUN 2008

Using Unsupervised Word Sense Disambiguation to Guess Verb Subjects on Untagged Corpora

PAULA CRISTINA VAZ
DAVID MARTINS DE MATOS
Spoken Language Laboratory, INESC-ID-Lisboa, Portugal

ABSTRACT

This article explores the use of subject lists extracted from an annotated corpus to find subject-verb pairs in untagged corpora. Our goal is to identify verb syntactic functions (subjects and direct objects) to characterize verb arguments. Since identifying syntactic functions on corpora using parsers is time-consuming, it is desirable to automate the annotation process of the syntactic functions without parsing the corpus. We present a method that uses a small annotated corpus to cluster sentences with synonymous verbs. We observe that verbs in the same cluster have the same list of nouns as subject in the test corpus, even though the specific pair subject/verb does not appear in the annotated corpus. The result shows that annotating the subject/verb pair using the subject lists extracted from the clusters is quicker than syntactically parsing the corpus.

1. INTRODUCTION

Our goal is to find nouns that can be used as subjects or objects using a small annotated manually-corrected corpus and without having to parse large amounts of corpora. Syntactically parsing usually means to lemmatize, disambiguate, chunk or find the parse tree, and, finally, connect the subject, object, and prepositional object with the predicate of the phrase or sentence. After the parsing process is complete, the corpus is ready to be searched for (subject, predicate), (object, predicate), or (preposition, predicate) pairs. Performing all these tasks is time-consuming. It becomes desirable to find a method of extracting syntactic functions without parsing the corpus.

A syntactically annotated corpus for the Portuguese language, Bosque Sintáctico (Afonso et al. 2002), is publicly available. To our knowledge, this corpus is the only one manually-corrected. This corpus is small (180k words) and the number of phrases available for each

LANGUAGE FORUM
VOL. 34, NO. 1, JAN-JUN 2008

Automatic Extraction of Case Frames for Chinese Verbs

XIAOHONG WU
WENLIANG CHEN
HITOSHI ISAHARA
National Institute of Information & Comm. Tech., Kyoto, Japan

ABSTRACT

Information on case frames of verbs is very useful for semantic and syntactic disambiguation needed in NLP, especially for parsing and MT where incorrect analysis or mismatching are frequently seen. However, extracting case frames for Chinese verbs is a difficult task as the Chinese language lacks obviously marked case frames. Prepositions could be alternatively considered as case markers; however, it is not as straightforward as we expect. Despite of this fact, there are still some structures in which the prepositions do play an important and special role in the syntactic constructions and can be considered as typical case markers, for example, the "ba" in ba-construction. In this paper we discuss the method that we employ to automatically extract case frames by using the prepositions as the key words as the first step for building a case frame dictionary. To get a good coverage as well as better and more reliable results we make use of large unlabelled corpora. We employ a dependency parser to parse the corpus, taking advantage of the dependency relationships provided by the parser to extract case frames. We lay focus on the kind of syntactic constructions in which the object is fronted by employing the prepositions and thus form a relatively fixed case frame. Or the prepositions are right after the verbs and also function as the object of the verbs. By defining strict rules using the context free grammar, we successfully extract the case frames of the verbs that can be used in such constructions and thereby get a list of verbs together with a list of nouns (its arguments) that can fill in such frames. This paper thus addresses the problems and solutions of such a task from a linguistic point of view, focusing on the methods which can facilitate the automatic extraction by the computer.

LANGUAGE FORUM
VOL. 34, NO. 1, JAN-JUN 2008

Mining Wikipedia as a Parallel and Comparable Corpus

JESÚS TOMÁS
JORDI BATALLER
AND FRANCISCO CASACUBERTA
Instituto Tecnológico de Informática

JAIME LLORET
Universidad Politécnica de Valencia

ABSTRACT

It is difficult to find parallel text corpora but for a few languages or for specific domains. Recently, collaborative edited multilingual projects, like Wikipedia, are becoming widespread. This paper studies the feasibility of using Wikipedia to obtain parallel corpora. Two types of articles can be used: parallel and comparable. We explore a distinct approach for each type of article. Parallel articles are identified by using – language independent – Web Mining techniques. For comparable articles, we use a classifier. It is based on a log-lineal combination of feature-functions, tuned with minimum error criterion. In order to validate our approaches, we report two experiments. First, a restricted domain corpus (pharmacology) in two great diffusion languages (English and Spanish) is obtained. In the second, the corpus is generated for a minority language (Catalan) and Spanish.

1. INTRODUCTION

Inductive methods are currently being considered with increasing interest in multilingual computational linguists. They have an important limitation: a parallel corpus is needed to train their models. For instance, a parallel corpus is required in tasks such as machine translation (Brown et al. 1993; Och & Ney 2000), cross-lingual information retrieval (Chen and Nie 2000) or lexical acquisition (Brown et al. 1991). There are few parallel corpora, for a few languages, and often in restricted domains. Currently, we can find large parallel

LANGUAGE FORUM
VOL. 34, NO. 1, JAN-JUN 2008

Cross-Language Transcription of Proper Names

EDWARD KLYSHINSKY
VADIM MAXIMOV
SERGEY YOLKEEN
M. V. Keldysh Institute of Applied Mathematics, Russian Academy of Sciences, Moscow, Russia

ABSTRACT

The paper is dedicated to the problem of automatic cross-language transcription of proper names. The correct written transcription of foreign proper names is a serious communication problem. It is especially important for legal translation of documents, data retrieval, postal processing and, in general, in all fields, where the accurate identification of places, persons and organizations is required. In order to formalize the process of transcription and reduce the number of errors, the automatic rule-based system of transcription has been developed. The system transcribes proper names between more than 20 languages, including non-European ones. The phonetic approach provides the easy integration of new languages in the system. The results of long-term collaboration of linguists and programmers had been generalized in the monograph.

1. INTRODUCTION

The accurate transfer of foreign proper names pronunciation by means of written transcription is an actual communicational problem, as it provides for the proper identification of places, people and organizations. The automatic solution of this problem will help to avoid certain mistakes in data retrieval in multi-language environment; then, it will facilitate the work of those specialists whose activity is connected with the transcription of foreign proper names: translators, postal operators and others who must process the foreign names in written form. The formalization of transcription rules will make it possible to enter the new quality level of cross-language transcription.

LANGUAGE FORUM
VOL. 34, NO. 1, JAN-JUN 2008

A Classification Approach to Automatic Evaluation of Machine Translation Based on Word Alignment

KATSUNORI KOTANI
Kansai Gaidai University,Osaka, Japan

TAKEHIKO YOSHIMI
Ryukoku University, Shiga, Japan

HITOSHI ISAHARA
National Institute of Information andComm. Tech, Kyoto, Japan

TAKESHI KUTSUMI
ICHIKO SATA
Sharp Corporation, Nara, Japan

ABSTRACT

Constructing a classifier that distinguishes machine translations from human translations is a promising approach to automatic evaluation of machine-translated sentences. Using this approach, we constructed a classifier based on word alignment distributions between source sentences and human/machine translations, using Support Vector Machines as machine learning algorithms. We found that word alignment distributions succeeded both in achieving a classification accuracy as high as 99.4% and in identifying the qualitative characteristics of machine translations, which greatly helps improve the quality of machine translations.

1. INTRODUCTION

Previous research proposed a classification approach to machine translation evaluation in which a machine translation system can be evaluated based on the extent to which machine-generated translations (MTs) are similar to human-generated translations (HTs) (Corston-Oliver, Gamon & Brockett 2001; Gamon, Aue & Smets 2005; Kulesza