Research Paper

Volume : 3 | Issue : 2 | February 2014 • ISSN No 2277 - 8179

Study on Analysis of Squential Complex Languages Through Machine (Technology) Learning

Engineering

KEYWORDS : Women, Mulberry sericulture, Farmers, Knowledge, Adoption

Pandya Pooja Piyushkumar

Monad University Delhi Hapur Road, Delhi.

INTRODUCTION

This study demonstrates that machine learning can be applied in different ways to automate the analysis of morphologically complex agglutinating languages. Issues explored are a general framework which combines morphological analysis and machine learning methods, evaluation metrics for assessing analysis algorithms, the generation of a morphological corpus for the target language Gujarati and second language as English as well as the presentation and experimental evaluation of novel algorithms for word decomposition and morpheme labeling.

Morphological analysis and its application

In natural language, words are often seen as the smallest atomic units. However, when examining them carefully, one can recognize reoccurring components. This study which is concerned with internal word structure, referred to as morphology [Booij, 2004, p.39], is called morphological analysis. At its heart, morphological analysis deals with the process of separating existing words into their constituents, called morphemes, and describing how words are built by combining these morphemes.

Traditionally, morphological analysis has been done in a manual fashion. It took linguistic experts several months to years in order to describe a single language. More recently, computer algorithms have been applied which can automate and therefore speed-up the process. These algorithms deploy techniques from machine learning to improve their performance with increasing numbers of words or analyzed word examples they have access to.

Morphology and its analysis play an important role in many natural language processing tasks. Automatic speech recognition (ASR) is concerned with the identification of spoken words and their transformation into text. Morphologically complex languages are especially challenging due to the combinatorial explosion of possible morpheme structures. As a consequence, a language description is likely to suffer from out-of-vocabulary (OOV) words – words that are not part of the vocabulary list. A solution could be either an increased vocabulary size with its implications on memory usage and processing time, or reverting to sub-word units such as morphemes. Arisoy et al. [2009] showed for Turkish that sub-word units improve acoustic and language modeling in comparison to sole word-based models.

Example 1.1.1. The words 'obtain', 'obtains', 'obtaining', 'obeying' and 'obeyed' are part of a known vocabulary. If a morphological system recognizes their grammatical similarity, it could come up with a rule [X] + [/0, '-s', '-ing', 'ed'] and a lexicon entry X = ['obey', 'occur']. Although 'obtained' and 'obeys' have not been seen before, they would be identified by the system.

In text-to-speech synthesis (TTS) a computational synthesizer produces the speech equivalent to an input text given to the system. As for ASR, TTS systems need to constrain the size of the word dictionary and have to deal with OOV words. This is done by performing morphological analysis such that words can be mapped to lexicon entries [Dutoit, 1997, pp. 77ff]. Furthermore, syntax together with morphology can help to resolve ambiguity when assigning grammatical categories and describing dependency relationships between successive words. Most importantly, however, there can be a relation between the phonetic transcription and the morphological structure of a word.

Example 1.1.2. The German words 'Illustration' (illustration) and 'Hauptstadt' (capital) both contain 'st'. Nevertheless, in the first word it is regularly pronounced as /st/ whereas it is /St/ in the second word since it resembles the root-initial there. Another example is 'th' in English. In 'lightheaded' it is part of two words which have been compounded, however, in 'anything' it is pronounced as /T/ since it occurs within a morpheme.1

The ability to produce words in a certain morphological context or derive the context not only improves speech technologies but also has its applications on spell checkers, information retrieval and machine translation.

A spell checker examines the spelling of words in a target language. For most languages it might suffice to have a large word dictionary. In morphologically complex languages like South African Zulu, millions of words can be derived by combining word stems or roots with prefixes and suffixes [Bosch and Eiselen, 2005]. Once again, for processing and memory efficiency, it is better to store a morpheme dictionary. A word to be checked would be morphologically analyzed and flagged as incorrect if it does not consist of allowable morphemes or morphemes are incorrectly combined.

Example 1.1.3. A spell-checker recognizes a mistake in the sentence 'He::::::: carried a box.' and proposes 'carried' since there exists a rule '-y + ed ! -ied' for the English past tense. This, however, implies that the system recognizes the word 'carried' as a verb to which the above rule applies.

Information retrieval faces various challenges, especially for non-English languages [Lazarinis et al., 2009]. Words in Semitic languages like Arabic, for instance, exhibit a phenomenon called root-and-pattern morphology. A pattern consists of vowels and place holders for consonants ('C'), e.g. 'CaCaC'. Word roots, on the other hand, are a skeleton of consonants, e.g. 'g-d-l'. Only when intertwined, the actual word 'gadal' (grow) is formed [Beesley, 2001; Booij, 2004, p.38].

Example 1.1.4. Moukdad [2004] demonstrated problematic characteristics of the Arabic morphology by using different variants of the same noun 'jamct' (university) – 'aljamct'

(the university), 'baljamct' (in the university), 'ljamcty' and 'wbaljamct' (and the university) in a retrieval experiment. For each word from the coverage of retrieved documents by four different search engines varied substantially.

A modern system for statistical machine translation (SMT) use as smallest units phrases rather than words; nevertheless, they have two major limitations [Nguyen and Shimazu, 2006]. Firstly, they do not utilize linguistic information. Phrases can have any order independent of their content. Secondly, translation systems for inflectional languages, i.e. languages that possess words in various forms, suffer from data sparsity since different forms are treated separately.

Example 1.1.5. A phrase-based system tries to translate 'Eu vou para casa' (I go home) from Portuguese to English but only knows the phrases 'Ela vai para casa' and 'She goes home'. Being able to recognize the exact forms of the verbs, 1st and 3rd Person Singular (3rd PS), and to transform one into the other is the key to solving this problem. This involves morphological and syntactic considerations.

Note: same examples are translated in the Guajarati language.

Under-resourced languages

Irrespective of the general applications of morphology and morphological analysis, the focus of this studies lies on under-resourced indigenous languages. Under-resourced means limited access to basic linguistic material like machine-readable text, dictionaries and grammars and also little or no availability of computational programs for morphological analysis. Other scarce resources are linguistic experts and, not to forget, funding.

Indigenous refers to languages which emerged in a specific region or place and were not brought from somewhere else, or to languages which were prevalent in a particular region before the emergence of other languages. Indigenous languages can be minority languages which find less attention in linguistic research than dominant world languages and might even face extinction.

For the purpose of this study, the indigenous of Gujarat state language second language as English and first language as Guajarati has been chosen as the target language and focus of research. It is one of the 22 official languages of INDIA, spoken by one fourth and understood by half of the population with a total number of around 10 million speakers.

Gujarati (Gujarati: ગુજરાતી Gujarātī) is an Indo-Aryan language, and part of the greater Indo European language family. It is derived from a language called Old Gujarati (1100–1500 AD) which is the ancestor language of the modern Gujarati and Rajasthani languages. It is native to the Indian state of Gujarat, where it is the chief language, and to the adjacent union territories of Daman and Diu and Dadra and Nagar Haveli.

There are about 65.5 million speakers of Gujarati worldwide, making it the 26th most spoken native language in the world. Along with Romany and Sindhi, it is among the most western of Indo-Aryan languages. Gujarati was the first language of Mohandas Karamchand Gandhi, the "Father of the Nation of India", and Sardar Vallabhbhai Patel, the "Iron Man of India." Other prominent personalities whose first language are or were Gujarati include Swami Dayananda Saraswati, Morarji Desai, Narsinh Mehta, Dhirubhai Ambani, and J. R. D. Tata. & Muhammad Ali Jinnah the "Father of the Nation of Pakistan"

Gujarati had been a solely vocal language until missionaries introduced a writing system based on the modern Indo-Aryan language evolved from Sanskrit alphabet in the eleventh century. In (1100–1500 AD), the first written document, Ramayana, Mahabharata, etc was published containing samples of the ethics which were translated by many authors and revised. The first printed book in Gujarati was written by a surgeon 200 years ago. Interestingly, it was written not by a Gujarati but an Englishman Dr Robert Drummond. In 1808, he wrote "Illustrations of the grammatical parts of the Guzerattee, Mahratta & English language," which is known as the first printed Gujarati book.

This and much more trivia about Gujarati publishing history is contained in Deepak Mehta's recently published book, Oganishmi Sadi Gujarati granth samrudhhi'. The book deals in interesting aspects of Gujarati books and men of letters. This well-researched work has colourful insights and inside stories on Gujarati books.

Not much is known about Drummond, except that he was appointed to the Bombay medical establishment in 1796. He was residency surgeon, Baroda and surgeon to the judge of appeal and circuit in Gujarat, and was struck off the rolls of the Bombay army on March 14, 1809, after getting lost at sea on his way home.

While in Gujarat, he learnt Gujarati. The 142-page book, separated into chapters, mainly contains glossary. Though the book was written in three languages, most of it was in Gujarati language.

The book also has sayings and information important from academic point of view. In preface, Drummond writes, "The following of the structures of two languages the most extensively spoken and written on the western shore of the Indian peninsula, were put together and printed, as the last and most essential trace of attachment that a person about to quit the theatre of them, himself, enabled to leave behind." Soon after book was published, Drummond died during his voyage to England. This first publication in Gujarati saw 467 advance bookings, majority by Englishmen.

In recent years, there have been several initiatives to improve resources for Gujarati by developing different computational approaches. Pretorius and Bosch [2003] implemented a prototype of a computational morphological analyzer for Gujarati based on the Xerox finite-state toolbox by Beesley and Karttunen [2003]. Later work dealt with over-generation of their system in terms of outputting or returning analyses for nonexistent Gujarati words. A bootstrapping approach to morphological analysis was pursued by Joubert et al. [2004]. A simple framework uses morpheme lists and rules which are learnt in a semi-automatic fashion. Botha and Barnard [2005] used commercial search engines and web crawlers for gathering Gujarati text corpora from the World Wide Web. Bosch and Eiselen [2005] presented a spell checker for Guajarati based on morphological analysis and regular expressions.

Nevertheless, no publicly available automatic analyzer and morphologically annotated corpus existed for Guajarati and English as the second language before this work. Furthermore, raw text has been mostly limited to the epics and other literature. An open-source corpus consisting of common words which have been morphologically analyzed would be valuable to linguists since it provides real world examples. Such a corpus could also be used to build and train automatic morphological analyzers which in turn could label large lists of new words. Incorporating new morphological analyzers into the above described systems for speech recognition and synthesis, spell checking, information retrieval and machine translation would have a positive impact on Gujarati as a written language, a language of instruction and a language in public life.

Machine learning background

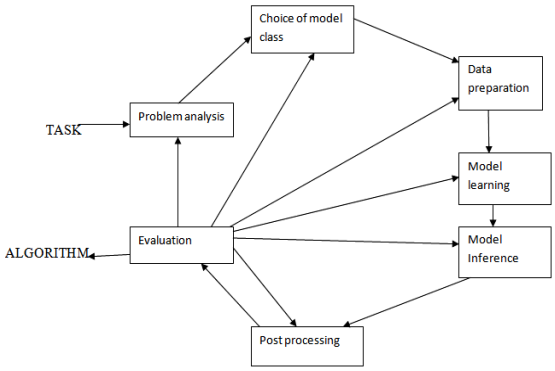
Having introduced the linguistic terminology necessary for morphological analysis, a brief overview of machine learning will be given. In general, machine learning is the systematic study of algorithms which improve their performance on a certain well-defined task with increasing experience. Tasks, among others, may be classification, clustering, ranking or regression. The experience refers to training examples which help the algorithm to advance its knowledge and therefore its performance on a given task. The performance is measured on new and unseen examples. Learning, in essence, is the ability to generalize from one set of examples establishing a hypothesis which can deal with new and previously unseen examples. Machine learning will be later applied to automate morphological analysis.

The machine learning work cycle

In this section activities will be described which are involved when developing and implementing machine learning algorithms. As an algorithm one generally understands a sequence of instructions for solving a certain task or problem. The algorithm bases on a certain model which was derived from a model class by model learning. Based on the model learnt a solution is constructed through model inference. Post-processing can either improve the model, e.g. the parameter estimates, or combine different models.

During the subsequent evaluation the performance of the algorithm is assessed and it is decided whether to return to earlier stages of the cycle for adoptions and improvements or not. The algorithm can be seen as the final product of the work cycle and the solution to a specific task or problem. An overview is given in Figure

Volume : 3 | Issue : 2 | February 2014 • ISSN No 2277 - 8179



REVIEW OF RELATED LITERATURE

Thomas G. Dietterich studied Statistical learning problems in many fields involve sequential data. It formalizes the principal learning tasks and describes the methods that have been developed within the machine learning re- search community for addressing these problems. These methods include sliding window methods, recurrent sliding windows, hidden Markov models, conditional random fields, and graph transformer networks. The paper also discusses some open research issues.

Konrad Rieck Studied Malicious software—so called malware—poses a major threat to the security of computer systems. The amount and diversity of its variants render classic security defenses ineffective, such that millions of hosts in the Internet are infected with malware in the form of computer viruses, Internet worms and Trojan horses. While obfuscation and polymorphism employed by malware largely impede detection at file level, the dynamic analysis of malware binaries during runtime provides an instrument for characterizing and defending against the threat of malicious software.

In this article, we propose a framework for the automatic analysis of malware behavior using machine learning. The framework allows for automatically identifying novel classes of malware with similar behavior (clustering) and assigning unknown malware to these discovered classes (classification). Based on both, clustering and classification, we propose an incremental approach for behavior-based analysis, capable of processing the behavior of thousands of malware binaries on a daily basis. The incremental analysis significantly reduces the run-time overhead of current analysis methods, while providing accurate discovery and discrimination of novel malware variants.

Direkt Profil (DP) is a system for grammatical profiling. It detects, annotates and displays grammatical constructs, both correct and incorrect, in freely-written texts by Swedish-speaking learners of French. It can also determine the learner's developmental stage, given a text with enough identifying attributes.

The scope of my work is the final step, the classification of the text as being typical for a certain stage, for which machine learning (ML) methods, more specifically

C4.5, LMT (Logistic Model Tree) and SVM (Support Vector Machine), have been applied.

This thesis is part of a longer-term research project, led by Jonas Granfeldt and Suzanne Schlyter at the Centre for languages and literature at Lund University.

The research project aims at increasing our knowledge regarding how Swedish speaking learners acquire proficiency in written French. During a three-year period, commencing in 2005, it is being financed by the Swedish Science Council. In my experiments with an early version (1.5.2) of the profiling system, precision and recall values of a ternary classified (basic/intermediate/native), based on support vector machines, have reached 70_83 %. The system has also been tested with C4.5- and logistic model tree-based classified, yielding similar (LMT) or slightly inferior (C4.5) results. Direkt Profil l 2.0 gives similar performance even for a quandary classified, and ternary classified precision and recall is somewhat improved as well (up to 88 %).

The Naive Byes method yields a small further overall precision/recall increase, and is much faster than SMO (SVM) on the experiment corpus. This project paves the way for further experiments with parameter selection and classifier performance.

METHODOLOGY

Research questions and objectives

In the previous sections, morphological analysis has been introduced, and reasons why and where it should be performed and important resources for morphologically complex indigenous languages. At this point, different areas this thesis will be dealing with are charted out by formulating a number of research questions. Initially, a careful consideration of terminology is needed which encompasses morphological analysis.

- What does morphological analysis comprise in terms of well-defined tasks, methods as well as input and output data?
- Are there different types of morphological analysis?
- What steps are required?
- Is there a certain structure which underpins all approaches for morphological analysis?
- What is involved in building a morphological analyzer in terms of algorithmic construction and subsequent morphological analysis?
- Finally, and most importantly, how can machine learning methods contribute to morphological analysis?
- The second area of interest is the generation of a corpus for the target language Gujarati and English.

Questions which arise are:

- How can the annotation process of a linguistic expert be assisted and speeded up?
- How can partial knowledge about the morphology of the target language be represented and used?
- Lastly, what kind of machine learning methods can be applied in order to automate annotation? The third issue explored is the construction of algorithms for morphological analysis. Having specified types of morphological analysis as part of the terminology, how can algorithms for certain types of morphological analysis are developed?
- What are suitable approaches?
- How can their performance be evaluated?

OBJECTIVES OF THE STUDY

- Summarizing the questions above, the thesis will address the following four objectives:
- The description of the general framework of morphological analysis.
- The description of all relevant processes for building, training, testing and evaluating an algorithm which performs morphological analysis.
- The examination of ways to evaluate morphological analysis results such that the assessment can be directly fed back into the development process for improvements.
- The compilation of a high-quality morphologically analyzed corpus for the indigenous language Gujarati and English as the second language in consultation with a linguistic expert.
- The construction of novel algorithms for morphological analysis of the indigenous language Gujarati and English along with the investigation of strengths and weaknesses of chosen approaches by performing experiments on annotated data.

Data preparation

When preparing data sets for machine learning experiments following issues have to be considered. An algorithm is trained by estimating its model parameters and then tested by carrying out inference based on the learnt model. Its performance also needs to be evaluated and possibly compared to other approaches. Although in most cases it is dealt with limited data, training and

evaluation should be carried out on different data sets since optimization on the same set can lead to over fitting training examples and any performance measured would be meaningless for new and unseen examples. Therefore, representative samples should be used for each stage of the work cycle.

For the above reasons, sample data will be divided into the following non-overlapping partitions. The training set is a subset which contains examples used during the learning phase of the model, e.g. for parameter estimation. The validation set, if necessary, contains examples which are consumed when tuning a model as it will be shown for calibration in later Section. The test set consists of the remaining disjoint examples which are deployed in model inference for the later evaluation of the algorithm's performance.

REFERENCE

B. Aarts and A. McMahon, editors. The Handbook of English Linguistics. Blackwell Publishing, 2006. 79 | | A. Abraham, G. Champion, J. C. Bryant, L. Grout, and J. L. Doehne. First Bible In Zulu. American Bible Society, New York, 1883. 4, 69 | | A. Albright and B. Hayes. Modeling English Past Tense Intuitions with Minimal Generalization. | Proceedings of the 6th Workshop of the ACL Special Interest Group in Computational Phonology (SIGPHON), pages 58–69, 2002. 44, 61, 65 | | A. Albright and B. Hayes. Rules vs. Analogy in English Past Tenses: a Computational/ Experimental Study. Cognition, 90:119–161, 2003. 61, 62, 65 | | S. R. Anderson. A-Morphous Morphology. Cambridge University Press, 1992. 13, 17 | | S. R. Anderson. Encyclopedia of Language & Linguistics, volume 5, chapter A Morphous | Morphology, pages 198–203. Elsevier, 2nd edition, 2006. 13, 17 | | S. Argamon, N. Akiva, A. Amir, and O. Kapah. Efficient Unsupervised Recursive Word Segmentation Using Minimum Description Length. Technical report, Illinois Institute of Technology, Dept. of Computer Science, Chicago, IL 60616, USA, Bar- Ilan University, Dept. of Computer Science, Ramat Gan 52900, ISRAEL, 2004. 52, 63 | | E. Arisoy, D. Can, S. Parlak, H. Sak, and M. Saraclar. Turkish Broadcast News Transcription and Retrieval. IEEE Transactions on Audio, Speech, and Language Processing, 17(5), july 2009. | | M. Aronoff. Word Formation in Generative Grammar. The MIT Press, 1976. 13 | | M. Aronoff and K. Fudeman. What is Morphology? (Fundamentals of Linguistics). Wiley-Blackwell, 2004. 12, 14 |