Natural Language Processing: A Review

Sethunya R Joseph[1],
Computer Science Department,
Botswana International University of Science and Technology,
Palapye, Botswana

Hlomani Hlomani[2]
Computer Science Department,
Botswana International University of Science and Technology,

Keletso Letsholo[3],
Computer Science Department,
Botswana International University of Science and Technology,
Palapye, Botswana

Freeson Kaniwa[4],
Computer Science Department,
Botswana International University of Science and Technology,
Palapye, Botswana

Kutlwano Sedimo[5]
Computer Science Department,
Botswana International University of Science and Technology,
Palapye, Botswana

ABSTRACT

Natural Language Processing (NLP) is a way of analyzing texts by computerized means. NLP involves gathering of knowledge on how human beings understand and use language. This is done in order to develop appropriate tools and techniques which could make computer systems understand and manipulate natural languages to perform various desired tasks. This paper reviews the literature on NLP. It also covers or gives a hint about the history of NLP. It is based on document analysis. This research paper could be beneficial to those who wish to study and learn about NLP.

Keywords: NLP, machine translation, machine learning, computational techniques, linguists

1. Introduction

Various researchers have explained Natural Language Processing (NLP) as an area of research and application that explores how computers can be used to understand and manipulate natural language text or speech to do useful things ([2]; [3]; [6]; [7]).

Liddy [1] defines NLP as a theoretically motivated range of computational techniques for analyzing and representing naturally occurring texts, at one or more levels of linguistic analysis for the purpose of achieving human-like language processing for a range of tasks or applications. The term NLP is normally used to describe the function of software or hardware components in a computer system which analyze or synthesize spoken or written language [14]. The 'natural' epithet is meant to distinguish human speech and writing from more formal languages, such as mathematical notations or programming languages, where vocabulary and syntax are comparatively restricted [14].

The research and development in NLP over the last sixty years as stated by Church and Rau [16] can be categorized into the following five areas:

- Natural Language Understanding
- Natural Language Generation
- Speech or Voice recognition
- Machine Translation
- Spelling Correction and Grammar Checking

The increase demands for softwares that process text of all kinds have tremendously been influenced by the advent of the Internet and World Wide Web. Over a decade, Internet publishing has become a common place activity for private individuals, commercial enterprises, and government organizations, as well as traditional media companies, and the medium of most of these communications and transactions is primarily natural language [14]. Various forms of keyword processing provide access to Web sites as well as organizational principles for retrieving, navigating and browsing web pages within those sites. Search engines and spam filters are now of everyday life and work well enough that their viability as products is not in question [14].

The language is more than transfer of information. Language is a set of resources to enable us to share meanings, but is not best thought of as a means for "encoding" meanings [16]. The foundations of NLP fall within a number of disciplines being: computer and information sciences, linguistics, mathematics, electrical and electronic engineering, psychology, artificial intelligence and robotics, etc. NLP applications comprise a number of fields of studies, such as natural language text processing and summarization,

machine translation, user interfaces, multilingual and cross language information retrieval, speech recognition, artificial intelligence and expert systems, and so on ([6]; [7]).

2. Scope and objective

Based on document analysis, this paper summarizes the information on NLP, the general overview, history, and previous works on NLP. It then considers applications of NLP. The challenges and failures of NLP together with current and future research of NLP are also discussed briefly in this paper. The research paper is intended to give an understating to researchers, scholarly peers and companies who wish to stay abreast with the NLP technologies and applications from the past, present and future.

3. Previous Works On NLP (Brief History)

NLP research dates back to the late 1940s with Machine translation (MT) being said to be the first computer-based application related to natural language. It was Weaver and Booth who started one of the earliest MT projects in 1946, on computer translation based on expertise in breaking enemy codes during World War II. However, a general agreement was made that, Weaver's memorandum of 1949 has brought the idea of MT to general notice and had inspired many projects. Weaver suggested using ideas from cryptography and information theory for language translation [1]; [29]. According to Liddy [1] earliest works in MT followed the basic view, that the only difference between languages was vested in their vocabularies and the permitted word orders. Hence systems which were made from this perspective basically used dictionary-lookup (for appropriate words for translation and reordering of the words after translation to fit the word-order rules of the target language). This was done without considering the lexical ambiguity inherent in natural language. This generated poor results and called for researchers to come up with a more sufficient theory of language. It was the Chomsky's 1957 publication [25] of the syntactic structures which introduced the idea of generative grammar [25], which gave the linguistic a better understanding of how they could help the machine translation. Subsequently, other NLP application areas began to emerge, such as speech recognition [1].

Since 1960 there have been some significant developments, both in production of prototype systems and in theoretical issues. This has mainly focused on the issue of how to represent meaning and developing computationally tractable solutions that the then-existing theories of grammar were not able to produce before 1960. Examples are: Chomsky's 1965 transformational model of linguistic [25]; case grammar of Fillmore [26], semantic networks of Quillian [27], and conceptual dependency theory of Schank, which explained syntactic anomalies, and provided semantic representations; Formalisms representation which included Wilks' preference semantics[28] and Kay's functional grammar; Augmented transition networks of Woods which extended the power of phrase-structure grammar by incorporating mechanisms from programming languages [10].

Besides theoretical development, many prototype systems have been developed. According to Liddy [1] these include: Weizenbaum's ELIZA [30] which was built to replicate the conversation between a psychologist and a patient, simple by permuting or echoing the user input; Winograd's SHRDLU simulation [8] of a robot that manipulated blocks on a tabletop which showed that natural language understanding was indeed possible for the computer [8], PARRY 's a theory of paranoia [31] in a system which used groups of keywords instead of single keywords and used synonyms if keywords were not found; LUNAR developed by Woods [9] as an interface system to a database that consisted of information about lunar rock samples using augmented transition network and procedural semantics ([1];[4]).

By the 1970's a substantial work was done on natural language generation, for example McKeown's discourse planner TEXT [32] and McDonald's response generator MUMMBLE [34] used rhetorical predicates to create declarative descriptions in short texts form (that is paragraphs) and TEXT's which generated comprehensible responses online. However, by the early 1980s, there was an increasing awareness of the limitations of isolated solutions to NLP problems and a general push towards applications that worked with language in a broad, real-world context. Since then to the present times, NLP has swiftly grown. This growth could be accredited to the advent of technologies such as: Internet; fast computers with increased memory; increased availability of large amounts of electronic text [1].

4. Natural Language Processing Overview

Given NLP's lineage, it is clear that many of its early theories and methods are derived from the field of linguistics [4]. A major shift was noticed in the early 1990s with the move to a reliance on empirical methodologies vs. the introspective generalizations that characterized the Chomsky era which held sway in theoretical linguistics. Liddy et al., [4] contends that, the focus in NLP shifted from what might be possible to do in a language and still have it be grammatically acceptable to what is actually observed to occur in naturally occurring text - that is, performance data. As more and larger corpora became available, empirical methods and evaluation rather than introspection-based methods and evaluation became the norm [4].

NLP researchers are now developing next generation NLP systems that deal reasonably well with general text and account for a good portion of the variability and ambiguity of a language. Statistical approaches thrived in dealing with many generic problems in computational linguistics such as part-of-speech identification, word sense disambiguation, etc., and have become standard throughout NLP [1].

Liddy [1]'s sentiments are also shared by Liddy et al., [4] that, the availability of larger, performance-oriented corpora supported the use of statistical (machine learning) methods, to learn the transformations that in previous approaches were performed by hand-built rules, eventually providing the empirical proof that statistical processing could accomplish some language analysis tasks at a level comparable to human performance. Liddy et al., [4] argue further that, at the center of this move lay the understanding that most of the work to be effected by language processing algorithms is too complex to be captured by rules constructed by human generalization, but rather require machine learning methods. According to Ringger et al., [8] the early statistical Part-Of-Speech tagging algorithms which were used in the early times, using Hidden Markov Models were said to achieve performance comparable to humans. In the course of the test sections of the Penn Treebank [35], and also on unobserved portions of the Brown Corpus [31], an up-to-date statistical parser was made known to perform more accurately than a broad-coverage rule-based parser [8]. Framing questions in the noisy channel model / information theory, with use of Probability Theory, Maximum Entropy, and Mutual Information, produced tangible advances in automatic capabilities [8].

The aforesaid transformations came in about because of the newly existing extensive electronic resources (e.g. the sizable corpora, such as the Brown corpus and other research programs) which were collected and distributed by the Linguistic Data Consortium. These were then followed by the lexical resources such as WordNet, which provided lexical-semantic knowledge bases (i.e. it enabled use of the semantic level of processing) and the Penn TreeBank (which provided gold standard syntactic resources that steered the development and testing of progressively rich algorithmic analysis tools [4].

A shift from a focus on closed domains of the earliest NLP research (from the 60s through the 80s) to open domains (e.g. newswire) has been made possible and supported by the increasing availability of realistically-sized resources coupled with machine learning methods. The flaring of the domains was further enabled by the availability of the broad ranging-textual resources of the web [4].

On the other hand, parallel with these moves towards use of more real world data, a realization was made that NLP researchers should evaluate their work on a larger scale, hence the introduction of empirically-based, blind evaluations across systems. These efforts led to the development of metrics such as BLEU and ROUGE that are integral to today's NLP research itself, of which they can be computed automatically and results fed back into the research [4]. Concomitant with these advances in statistical capabilities, but moving at a slower pace, was the demonstration that higher levels of human language analysis are amenable to NLP. The lower levels (morphological, lexical, and syntactic) deal with smaller units of analysis and are considered to be more rule-oriented and therefore more amenable to statistical analysis, while the higher levels (with semantics as a middle level, and discourse and pragmatics as the higher levels) admit of more free choice and variability in usage. This is to mean that, these levels allow more variation, with more exceptions, and perhaps less regularity (e.g. Rhetorical Structure Theory in NLP by Mann & Thompson [10], demonstrated that even much larger units of analysis (e.g., treatises, instructional guides, etc.) are amenable to computational analysis [4].

Wiebe et al., [13] state that, in information extraction, increasingly complex phenomena such as subjectivity and opinion are identified automatically. Charniak et al., [11] and Quirk et al., [12] point out that with the most recent machine translation results, syntax based MT outperforms surface-level word and phrase replacement systems. These developments have resulted in the realization that NLP, by the blending of statistical and symbolic methods, together with lexical resources such as WordNet, and syntactic and semantic resources such as Prop Bank, plus the availability of large scale corpora on which to test and evaluate approaches, is gaining ground on the goal of realistic comprehension and production of human-like language understanding [4].

5. Applications of NLP

According to Church and Rau [16], in recent years, the natural language text interpretation and processing technologies have also gained an increasing level of sophistication. For example, generic engines are now available which can deliver semantic representations for sentences, or deliver sentences from representations. It is now possible to build very-targeted systems for specific purposes, for example, finding index terms in open text, and also the ability to judge what level of syntax analysis is appropriate. NLP technologies are becoming extremely important in the creation of user-friendly decision-support systems for everyday non-expert users, particularly in the areas of knowledge acquisition, information retrieval and language translation [16].

NLP technology has progressively increased. It can be noted that this has happened because of the following reasons: The web has provided researchers with readily accessible corpus of electronic document on scale that is unprecedented ; Academia has replaced a new emphasis upon empirical approaches to language processing that rely more heavily upon corpus statistics than linguist theory and Modern networked machines are capable of processing millions of documents and performing the billions of calculations to build statical profiles of large corpora[14].

Massive quantities of text are becoming available in electronic form, ranging from published documents such as electronic dictionaries, encyclopedias, libraries and archives for information retrieval services, private databases, personal email and faxes [15]. Online information services are reaching mainstream computer users. With media attention reach time, hardly a day goes by without a new article on the national information infrastructure, digital libraries, networked services, digital convergence or intelligent agents. This attention is moving NLP along the critical path for all kinds of novel applications [15] see Figure 1.

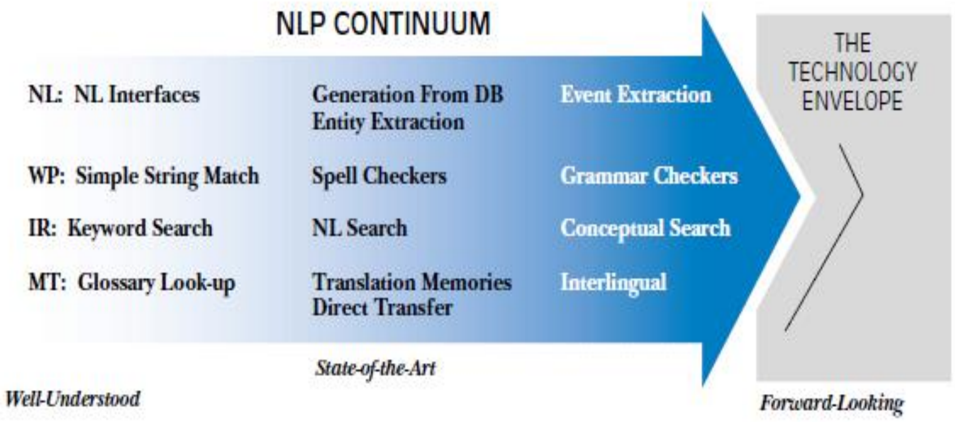

Figure 1: A diagram showing the NLP continuum (adopted from Church and Rau [16])

Figure 1 shows a number of technologies ranging from well-understood technologies, such as string matching, to more forward looking technologies such as grammar checkers, conceptual search, event extraction, interlingual and so on.

There are so many examples of the aforementioned technologies, However this paper does not provide an exhaustive list (not all of them will be elaborated in this paper), examples of some applications products of NLP as stated by Church and Rau [16] which are given in this paper includes: Word Processing and Desktop Publishing, WordPerfect (Novell) which offers Grammatik 6, which checks for grammatical errors and attempts to fix them. Microsoft has demonstrated a considerable long-term commitment to improving the technology by hiring a research group, for significant contributions to grammar checking [24]. Other examples of the application products include: finite-state automata (a practical set of algorithms that efficiently represent functions from strings to strings); Transducer techniques (facilitates the retrieval of answers to service repair questions from a text database of repair manual); information retrieval products such as Xerox XSoft's Visual Recall, and OCR products such as Xerox Imaging Systems' Textbridge; lexical products through the Desktop Document Systems (DDS) division (which uses the technology to create a range of multilingual components that can be embedded in information retrieval, translation, and other document management applications);spelling checkers (e.g. The Xerox Memory-Writer typewriter). Kaplan and Kay have been working on these issues for over a decade at Xerox PARC [22].

The applications (such as the spell checker for the Xerox MemoryWriter typewriter) were the basis for a start-up company called Microlytics, formed in 1985 and later (1987) merged into the publically traded Selectronics Corp [16]. Through Microlytics, the original Kaplan and Kay algorithms found their way into spelling checkers and thesaurii, such as those included in popular systems such as Micropro, Claris, MacWrite II, Microsoft Word 4 (the thesaurus), Symantec, and WordFinder software sold to the PC and Apple Macintosh user community.

The hand-held language-related device market is now dominated by such companies as Casio, Seiko, Fuji, Xerox, Eurotronics, Franklin, Sharp and other primarily Asian manufacturers [16]. Word processing and information management were previously cited as two of the better examples of commercial opportunities for natural language processing [16]. The importance of information management is beginning to be appreciated as vast quantities of text become available in electronic form: digital libraries, it wasn't all that long ago that the researchers referred to the Brown Corpus as a "large'' corpus [21]; [31]. However, Dialog, Westlaw, Lexis-Nexis and other major vendors of online information services are archiving hundreds of megabytes per night, the equivalent of one Brown Corpus per hour [16]. Another significant recent development on NLP applications is the development of a speller checker for the Tswana native language, which has been designed to work on the Mozilla Firefox software and Libre-Office application [19].

6. Challenges and failures

Church and Rau [16] points out that even though we should know better, it is so appealing to fantasize about intelligent computers that understand human communication, that hyperbole is practically unavoidable. Sometimes these practices work out for the best. Symantec, for example, a highly successful vendor of software tools for the PC, started with a product called Q&A, an NLP program for querying a database. The Q&A was successful because of its unique packaging of AI/NLP with a good simple database facility. Neither would have been successful in isolation. The AI/NLP generated initial sales, but the real value was in the database. People bought the product because they were intrigued with the AI/NLP technology, but most users ended up turning off the AI/NLP features [20]. But all too often excessive optimism results in a manic-like cycle of euphoric activity followed by severe depression. In 1954, Georgetown University demonstrated what would now be called a "toy'' system. It was designed to translate a small corpus of approximately 50 Russian sentences into English. Little if any attempt was made to generalize to sentences beyond the tiny test corpus [16]; [29].

The limitations of today's practical language processing technology have been summarized by Bobrow and Weischedel [18] as follows:

1. Current systems have limited discourse capabilities that are almost ex-clusively handcrafted. Thus current systems are limited to viewing interaction, translation, and writing text as processing a sequence of either isolated sentences or loosely related paragraphs. Consequently, the user must adapt to such limited discourse.

2. Domains must be narrow enough so that the constraints on the relevant semantic concepts and relations can be expressed using current knowledge presentation techniques, i.e., primarily in terms of types and sorts. Processing may be viewed abstractly as the application of recursive tree re-writing rules, including filtering out trees not matching a certain pattern.

3. Handcrafting is necessary, particularly in the grammatical components of systems (the component technology that exhibits least dependence on the application domain). Lexicons and axiomatizations of critical facts must be developed for each domain, and these remain time-consuming tasks.

4. The user must still adapt to the machine, but, as the products testify, the user can do so effectively.

7. Current and Future progress of NLP

Some of the active researches on NLP phenomena include the Syntactic phenomena: those that pertain to the structure of a sentence and the order of words in the sentence, based on the grammatical classes of words rather than their meaning (e.g. discriminative models for scoring parses, coarse to fine efficient approximate parsing, dependency grammar); Machine translation (e.g. models and algorithms, low- resource and morphological complex language); Semantic phenomena : those that pertain to the meeting of a sentence relatively independent of the context in which the language occurs(e.g. sentiment analysis, summarization, information extraction ,slot-filling, discourse analysis, textual entailment);Pragmatic phenomena such as Speech: those that relate the meaning of a sentence to the context in which it occurs. This context can be linguistic (such as the previous text or dialogue) or, non-linguistic (such as knowledge about person who produced the language, about goals of the communication, about the objects in the current visual field, etc. (e.g. language modelling-syntax and semantics, models of acoustics, pronunciation) [17]; [18]. Speech recognition and information retrieval have finally gone commercial and there is a ton of text and speech on the Internet, cell phones, etc. It is now clear that studies regarding anything about a language are possible, e.g. formalizing some insights e.g. discrete knowledge (what is possible) and continuous knowledge (what is likey); studying the formalism mathematically; developing and implementing algorithms and testing on real data. The current and on-going future changes or improvements which need to be done to NLP are: to add features to existing interfaces, back end processing should be fully implemented (e.g. information extraction and normalization to build databases. Another anticipated improvement is of having hand held devices with translators and personal conversation recorder with topical searches [17].

8. Conclusions

As a computerized approach of analyzing text, NLP is continually striving forward. Researchers are continually trying to gather knowledge on how human beings understand and use various languages. This aid in the development of appropriate tools and techniques which make computer systems understand and manipulate natural languages to perform the various tasks. Technologies, such as string matching, keyword search, glossary lookup are now on the past as, to more forward looking technologies such as grammar checkers, conceptual search, event extraction, interlingual on going and striving forward.

References

[1] E.D. Liddy, Natural Language Processing, 2001.

[2] N. Kaur1, V. Pushe and R Kaur,"Natural Language Processing Interface for Synonym", International Journal of Computer Science and Mobile Computing, Vol.3 Issue.7, July- 2014, pp. 638-642 ,ISSN 2320–088X.

[3] S. Vijayarani1, J. Ilamathi and Nithya, "Preprocessing Techniques for Text Mining - An Overview", International Journal of Computer Science & Communication Networks, Vol.5, issue.1, pp. 7-16 7 ISSN: 2249-5789

[4]L.Liddy, E. Hovy, J.Lin, J.Prager, D. Radev, L.Vanderwende, R.Weischedel, "Natural Language Processing", This report is one of five reports that were based on the MINDS workshops.

[5] G.Chowdhury, "Natural language processing", Annual Review of Information Science and Technology, 2003, 37. pp. 51-89, ISSN 0066-4200.

[6] S. Jusoh and H.M. Alfawareh, "Natural language interface for online sales", in Proceedings of the International Conference on Intelligent and Advanced System (ICIAS2007),Malaysia: IEEE, November 2007, pp. 224-228

[7] E.K. Ringger, R.C. Moore, E. Charniak, L. Vanderwende, and H Suzuki, "Using the Penn Treebank to Evaluate Non-Treebank Parsers", In Proceedings of the 2004 Language Resources and Evaluation Conference (LREC), 2004, Lisbon, Portugal.

[8] T. Winograd, Procedures as a Representation for Data in a Computer Program for Understanding Natural Language, 1971, MIT-AI-TR-235

[9] W. A. Woods, "Transition Network Grammars for Natural Language Analysis", Communications of the ACM 13:10, 1970.

[10] W.C. Mann & S. Thompson, "Rhetorical Structure Theory: Toward a Functional Theory of Text Organization", 1988. Text 8 (3). Pp. 243-281.

[11] E. Charniak, K. Knight, and K.Yamada, "Syntax-based Language Models for Statistical Machine Translation". *In Proceedings of MT Summit IX*, 2003.

[12] C. Quirk, A. Menezes and C. Cherry, "Dependency Treelet Translation: Syntactically Informed Phrasal SMT". *In Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics*, Ann Arbor, Michigan, 2005.

[13] J. Wiebe, E. Breck, C. Buckley, C. Cardie, P. Davis, B. Fraser, D. Litman, D. Pierce, E. Riloff, T. Wilson, D. Day, and M. Maybury, "Recognizing and Organizing Opinions Expressed in the World Press". *In Proceedings of 2003 AAAI Spring Symposium on New Directions in Question Answering,* 2003.

[14] P. Jackson and I. Moulinier,"Natural Language Processing for Online Applications": Cambridge University press, New York.2012, page 7-9.

[15] R. Bose. "Natural language processing: Current state and future directions". *International Journal of the Computer, the Internet and Management* Vol. 12#1 (January – April, 2004) pp. 1 – 11.

[16] K. W Church and L.F Rau," Commercial applications of Natural Language Processing". *Communication of the ACM*, vol 38, No. 11,November 1995

[17] J. Eisner. Current and future NLP research.

[18] R. J Bobrow and R.M. Weischedel, "Challenges in Natural Language Processing," Cambridge University press, New York.1993

[19] D. Bailey et al.,"Tswana Spell Checker", available online: https://addspns.mozilla.org/.../fi.../addon/tswana-spell-checker.last accessed 11/14/2015.

[20] W. Frakes and R., Baeza-Yates, Eds," Information Retrieval: Data Structures & Algorithms". Prentice Hall, Englewood Cliffs, NJ, 1992.

[22] W. Francis and H .K. Houghton Mifflin, Brown University, 1982.

[23] R.M. Kaplan and M. Kay, "Regular models of phonological rule systems", *American J. Computational Linguistics*, 1994.

[24] K., Jensen, G. Heidorn, and S. Richardson, "Natural Languag Processing: The PLNLP Approach," Kluwer Academic Publisher, Boston, 1993

[25] N. Chomsky.Syntatic structures.The Hague: Mouton & Co. Reprinetd 1978,Peter Lang Publisjing

[26] C. J. Fillmore," The Case for Case". In Bach and Harms (Ed.): *Universals in Linguistic Theory*. New York: Holt, Rinehart, and Winston, 1-88, 1968.

[27] R. Quillian, "A notation for representing conceptual information: An application to semantics and mechanical English para- phrasing", SP-1395, System Development Corporation, Santa Monica, 1963.

[28] Y. Wilks, Preference Semantics",1973.

[29] J. Hutchins.,"The history of Machine translation in a nutshell".Available[online] : http://ourworld.compuserve.com/homepages/WJHutchins,Revised November 2005.Retrieved 11/13/2015.

[30] J. Weizenbaum, "ELIZA-A Computer Program For the Study of Natural Language Communication Between Man And Machine", *Communications of the ACM Vol.***9.,** No.1pp:36-45,January 1966), doi:10.1145/365153.365168

[31] V. Cerf, "PARRY encounters the DOCTOR ,"IETF. RFC 439,"21 January 1972.

[32] K. R McKeown., "Discourse Strategies for generating Natural Language Text". Artificial Intelligence.

[33] W.N. Francis and H. Kucera,"The Brown Corpus Manual",1964,Available[online]: http://clu.uni.no/icame/brown/bcm.html.Accessed: 13/11/2015

[34] R. Rubinoff, "Adapting MUMBLE: experience with natural Language generation", Published in Proceeding HLt'86 Proceedings of the workshop on Strategic computing natural language pp. 200-211.

[35] A. Taylor, "The Penn Tree bank: Overview," Available at http://www.ldc.upenn.edu