UNIT 3 NATURAL LANGUAGE PROCESSING

CONTENTS

1.0 INTRODUCTION

Natural language processing (NLP) is a field of computer science and linguistics concerned with the interactions between computers and human (natural) languages; it began as a branch of artificial intelligence. In theory, natural language processing is a very attractive method of human–computer interaction. Natural language understanding is sometimes referred to as an AI-complete problem because it seems to require extensive knowledge about the outside world and the ability to manipulate it.

An automated online assistant providing customer service on a web page is an example of an application where natural language processing is a major comp1nt.

2.0 OBJECTIVES

At the end of this unit, you should be able to:

Describe the history of natural language processing
List major tasks in NLP
Mention different types of evaluation of NPL.

3.0 MAIN CONTENT

3.1 History of natural language processing

The history of NLP generally starts in the 1950s, although work can be found from earlier periods. In 1950, Alan Turing published his famous article "Computing Machinery and Intelligence" which proposed what is now called the Turing test as a criterion of intelligence. This criterion depends on the ability of a computer program to impersonate a human in a real-time written conversation with a human judge, sufficiently well that the judge is unable to distinguish reliably — on the basis of the conversational content alone — between the program and a real human. The Georgetown experiment in 1954 involved fully automatic translation of more than sixty Russian sentences into English. The authors claimed that within three or five years, machine translation would be a solved problem. However, real progress was much slower, and after the ALPAC report in 1966, which found that ten years long research had failed to fulfil the expectations, funding for machine translation was dramatically reduced. Little further research in machine translation was conducted until the late 1980s, when the first statistical machine translation systems were developed.

Some notably successful NLP systems developed in the 1960s were SHRDLU, a natural language system working in restricted "blocks worlds" with restricted vocabularies, and ELIZA, a simulation of a Rogerian psychotherapist, written by Joseph Weizenbaum between 1964 to 1966. Using almost no information about human thought or emotion, ELIZA sometimes provided a startlingly human-like interaction. When the "patient" exceeded the very small knowledge base, ELIZA might provide a generic response, for example, responding to "My head hurts" with "Why do you say your head hurts?”

During the 70's many programmers began to write 'conceptual ontologies', which structured real-world information into computer-understandable data. Examples are MARGIE (Schank, 1975), SAM (Cullingford, 1978), PAM (Wilensky, 1978), TaleSpin (Meehan, 1976), QUALM (Lehnert, 1977), Politics (Carbonell, 1979), and Plot Units (Lehnert 1981). During this time, many chatterbots were written including PARRY, Racter, and Jabberwacky.

Up to the 1980s, most NLP systems were based on complex sets of hand-written rules. Starting in the late 1980s, however, there was a revolution in NLP with the introduction of machine learning algorithms for language processing. This was due both to the steady increase in computational power resulting from Moore's Law and the gradual lessening of the dominance of Chomskyan theories of linguistics (e.g. transformational grammar), whose theoretical underpinnings discouraged the sort of corpus linguistics that underlies the machine-learning approach to language processing. Some of the earliest-used machine learning algorithms, such as decision trees, produced systems of hard if-then rules similar to existing hand-written rules. Increasingly, however, research has focused on statistical models, which make soft, probabilistic decisions based on attaching real-valued weights to the features making up the input data. Such models are generally more robust when given unfamiliar input, especially input that contains errors (as is very common for real-world data), and produce more reliable results when integrated into a larger system comprising multiple subtasks.

Many of the notable early successes occurred in the field of machine translation, due especially to work at IBM Research, where successively more complicated statistical models were developed. These systems were able to take advantage of existing multilingual textual corpora that had been produced by the Parliament of Canada and the European Union as a result of laws calling for the translation of all governmental proceedings into all official languages of the corresponding systems of government. However, most other systems depended on corpora specifically developed for the tasks implemented by these systems, which was (and often continues to be) a major limitation in the success of these systems. As a result, a great deal of research has gone into methods of more effectively learning from limited amounts of data.

Recent research has increasingly focused on unsupervised and semi-supervised learning algorithms. Such algorithms are able to learn from data that has not been hand-annotated with the desired answers, or using a combination of annotated and non-annotated data. Generally, this task is much more difficult than supervised learning, and typically produces less accurate results for a given amount of input data. However, there is an enormous amount of non-annotated data available (including, among other things, the entire content of the World Wide Web), which can often make up for the inferior results.

3.2 NLP Using Machine Learning

As described above, modern approaches to natural language processing (NLP) are grounded in machine learning. The paradigm of machine learning is different from that of most prior attempts at language processing. Prior implementations of language-processing tasks typically involved the direct hand coding of large sets of rules. The machine-learning paradigm calls instead for using general learning algorithms — often, although not always, grounded in statistical inference — to automatically learn such rules through the analysis of large corpora of typical real-world examples.

A corpus (plural, "corpora") is a set of documents (or sometimes, individual sentences) that have been hand-annotated with the correct values to be learned.

As an example, consider the task of part of speech tagging, i.e. determining the correct part of speech of each word in a given sentence, typically one that has never been seen before. A typical machine-learning-based implementation of a part of speech tagger proceeds in two steps, a training step and an evaluation step. The first step — the training step — makes use of a corpus of training data, which consists of a large number of sentences, each of which has the correct part of speech attached to each word. (An example of such a corpus in common use is the Penn Treebank. This includes (among other things) a set of 500 texts from the Brown Corpus, containing examples of various genres of text, and 2500 articles from the Wall Street Journal.) This corpus is analyzed and a learning model is generated from it, consisting of automatically created rules for determining the part of speech for a word in a sentence, typically based on the nature of the word in question, the nature of surrounding words, and the most likely part of speech for those surrounding words. The model that is generated is typically the best model that can be found that simultaneously meets two conflicting objectives: To perform as well as possible on the training data, and to be as simple as possible (so that the model avoids over fitting the training data, i.e. so that it generalizes as well as possible to new data rather than only succeeding on sentences that have already been seen). In the second step (the evaluation step), the model that has been learned is used to process new sentences. An important part of the development of any learning algorithm is testing the model that has been learned on new, previously unseen data. It is critical that the data used for testing is not the same as the data used for training; otherwise, the testing accuracy will be unrealistically high.

Many different classes of machine learning algorithms have been applied to NLP tasks. In common to all of these algorithms is that they take as input a large set of "features" that are generated from the input data. As an example, for a part-of-speech tagger, typical features might be the identity of the word being processed, the identity of the words immediately to the left and right, the part-of-speech tag of the word to the left, and whether the word being considered or its immediate neighbors are content words or function words. The algorithms differ, however, in the nature of the rules generated. Some of the earliest-used algorithms, such as decision trees, produced systems of hard if-then rules similar to the systems of hand-written rules that were then common.

Increasingly, however, research has focused on statistical models, which make soft, probabilistic decisions based on attaching real-valued weights to each input feature. Such models have the advantage that they can express the relative certainty of many different possible answers rather than only one, producing more reliable results when such a model is included as a component of a larger system. In addition, models that make soft decisions are generally more robust when given unfamiliar input, especially input that contains errors (as is very common for real-world data).

Systems based on machine-learning algorithms have many advantages over hand-produced rules:

> The learning procedures used during machine learning automatically focus on the most common cases, whereas when writing rules by hand it is often not obvious at all where the effort should be directed.
>
> Automatic learning procedures can make use of statistical inference algorithms to produce models that are robust to unfamiliar input (e.g. containing words or structures that have not been seen before) and to erroneous input (e.g. with misspelled words or words accidentally omitted). Generally, handling such input gracefully with hand-written rules — or more generally, creating systems of hand-written rules that make soft decisions — is extremely difficult and error-prone.
>
> Systems based on automatically learning the rules can be made more accurate simply by supplying more input data. However, systems based on hand-written rules can only be made more accurate by increasing the complexity of the rules, which is a much more difficult task. In particular, there is a limit to the complexity of systems based on hand-crafted rules, beyond which the systems become more and more unmanageable. However, creating more data to input to machine-learning systems simply requires a corresponding increase in the number of man-hours worked, generally without significant increases in the complexity of the annotation process.

3.3 Major tasks in NLP

The following is a list of some of the most commonly researched tasks in NLP. Note that some of these tasks have direct real-world applications, while others more commonly serve as subtasks that are used to aid in solving larger tasks. What distinguishes these tasks from other potential and actual NLP tasks is not only the volume of research devoted to them but the fact that for each one there is typically a well-

Automatic summarization: Produce a readable summary of a chunk of text. Often used to provide summaries of text of a known type, such as articles in the financial section of a newspaper.
Co reference resolution: Given a sentence or larger chunk of text, determine which words ("mentions") refer to the same objects ("entities"). Anaphora resolution is a specific example of this task, and is specifically concerned with matching up pronouns with the nouns or names that they refer to. The more general task of co reference resolution also includes identify so-called "bridging relationships" involving referring expressions. For example, in a sentence such as "He entered John's house through the front door", "the front door" is a referring expression and the bridging relationship to be identified is the fact that the door being referred to is the front door of John's house (rather than of some other structure that might also be referred to).
Discourse analysis: This rubric includes a number of related tasks. One task is identifying the discourse structure of connected text, i.e. the nature of the discourse relationships between sentences (e.g. elaboration, explanation, contrast). Another possible task is recognizing and classifying the speech acts in a chunk of text (e.g. yes-no question, content question, statement, assertion, etc.).
Machine translation: Automatically translate text from one human language to another. This is one of the most difficult problems, and is a member of a class of problems colloquially termed "AI-complete", i.e. requiring all of the different types of knowledge that humans possess (grammar, semantics, facts about the real world, etc.) in order to solve properly.
Morphological segmentation: Separate words into individual morphemes and identify the class of the morphemes. The difficulty of this task depends greatly on the complexity of the morphology (i.e. the structure of words) of the language being considered. English has fairly simple morphology, especially inflectional morphology, and thus it is often possible to ignore this task entirely and simply model all possible forms of a word (e.g. "open, opens, opened, and opening") as separate words. In languages such as Turkish, however, such an approach is not possible, as each dictionary entry has thousands of possible word forms.

Named entity recognition (NER): Given a stream of text, determine which items in the text map to proper names, such as people or places, and what the type of each such name is (e.g. person, location, organization). Note that, although capitalization can aid in recognizing named entities in languages such as English, this information cannot aid in determining the type of named entity, and in any case is often inaccurate or insufficient. For example, the first word of a sentence is also capitalized, and named entities often span several words, only some of which are capitalized. Furthermore, many other languages in non-Western scripts (e.g. Chinese or Arabic) do not have any capitalization at all, and even languages with capitalization may not consistently use it to distinguish names. For example, German capitalizes all nouns, regardless of whether they refer to names, and French and Spanish do not capitalize names that serve as adjectives.

Natural language generation: Convert information from computer databases into readable human language.

Natural language understanding: Convert chunks of text into more formal representations such as first-order logic structures that are easier for computer programs to manipulate. Natural language understanding involves the identification of the intended semantic from the multiple possible semantics which can be derived from a natural language expression which usually takes the form of organized notations of natural languages concepts. Introduction and creation of language metamodel and ontology are efficient however empirical solutions. An explicit formalization of natural languages semantics without confusions with implicit assumptions such as closed world assumption (CWA) vs. open world assumption, or subjective Yes/No vs. objective True/False is expected for the construction of a basis of semantics formalization.

Optical character recognition (OCR): Given an image representing printed text, determine the corresponding text.

Part-of-speech tagging: Given a sentence, determine the part of speech for each word. Many words, especially common ones, can serve as multiple parts of speech. For example, "book" can be a noun ("the book on the table") or verb ("to book a flight"); "set" can be a noun, verb or adjective; and "out" can be any of at least five different parts of speech. Note that some languages have more such ambiguity than others. Languages with little inflectional morphology, such as English are particularly prone to such ambiguity. Chinese is prone to such ambiguity because it is a tonal language during verbalization. Such inflection is not readily conveyed via the entities employed within the orthography to convey intended meaning.

Parsing: Determine the parse tree (grammatical analysis) of a given sentence. The grammar for natural languages is ambiguous and typical sentences have multiple possible analyses. In fact, perhaps surprisingly, for a typical sentence there may be thousands of potential parses (most of which will seem completely nonsensical to a human).

Question answering: Given a human-language question, determine its answer. Typical questions have a specific right answer (such as "What is the capital of Canada?"), but sometimes open-ended questions are also considered (such as "What is the meaning of life?").

Relationship extraction: Given a chunk of text, identify the relationships among named entities (e.g. who is the wife of whom).

Sentence breaking (also known as sentence boundary disambiguation): Given a chunk of text, find the sentence boundaries. Sentence boundaries are often marked by periods or other punctuation marks, but these same characters can serve other purposes (e.g. marking abbreviations).

Sentiment analysis: Extract subjective information usually from a set of documents, often using online reviews to determine "polarity" about specific objects. It is especially useful for identifying trends of public opinion in the social media, for the purpose of marketing.

Speech recognition: Given a sound clip of a person or people speaking, determine the textual representation of the speech. This is the opposite of text to speech and is one of the extremely difficult problems colloquially termed "AI-complete" (see above). In natural speech there are hardly any pauses between successive words, and thus speech segmentation is a necessary subtask of speech recognition (see below). Note also that in most spoken languages, the sounds representing successive letters blend into each other in a process termed coarticulation, so the conversion of the analog signal to discrete characters can be a very difficult process.

Speech segmentation: Given a sound clip of a person or people speaking, separate it into words. A subtask of speech recognition and typically grouped with it.

Topic segmentation and recognition: Given a chunk of text, separate it into segments each of which is devoted to a topic, and identify the topic of the segment.

Word segmentation: Separate a chunk of continuous text into separate words. For a language like English, this is fairly trivial, since words are usually separated by spaces. However, some written languages like Chinese, Japanese and Thai do not mark word boundaries in such a fashion, and in those languages text segmentation is a significant task requiring knowledge of the vocabulary and morphology of words in the language.

Word sense disambiguation: Many words have more than one meaning; we have to select the meaning which makes the most sense in context. For this problem, we are typically given a list of words and associated word senses, e.g. from a dictionary or from an online resource such as WordNet.

In some cases, sets of related tasks are grouped into subfields of NLP that are often considered separately from NLP as a whole. Examples include:

Information retrieval (IR): This is concerned with storing, searching and retrieving information. It is a separate field within computer science (closer to databases), but IR relies on some NLP methods (for example, stemming). Some current research and applications seek to bridge the gap between IR and NLP.

Information extraction (IE): This is concerned in general with the extraction of semantic information from text. This covers tasks such as named entity recognition, coreference resolution, relationship extraction, etc.

Speech processing: This covers speech recognition, text-to-speech and related tasks.

Other tasks include:

Stemming
Text simplification
Text-to-speech
Text-proofing
Natural language search
Query expansion
Truecasing

3.4 Statistical Natural Language Processing

Statistical natural-language processing uses stochastic, probabilistic and statistical methods to resolve some of the difficulties discussed above, especially those which arise because longer sentences are highly ambiguous when processed with realistic grammars, yielding thousands or millions of possible analyses. Methods for disambiguation often involve the use of corpora and Markov models. Statistical NLP comprises all quantitative approaches to automated language processing, including probabilistic modeling, information theory, and linear algebra. The technology for statistical NLP comes mainly from machine learning and data mining, both of which are fields of artificial intelligence that involve learning from data.

3.5 Evaluation of natural language processing

3.5.1 Objectives

The goal of NLP evaluation is to measure one or more qualities of an algorithm or a system, in order to determine whether (or to what extent) the system answers the goals of its designers, or meets the needs of its users. Research in NLP evaluation has received considerable attention, because the definition of proper evaluation criteria is one way to specify precisely an NLP problem, going thus beyond the vagueness of tasks defined only as language understanding or language generation. A precise set of evaluation criteria, which includes mainly evaluation data and evaluation metrics, enables several teams to compare their solutions to a given NLP problem.

3.5.2 Short history of evaluation in NLP

The first evaluation campaign on written texts seems to be a campaign dedicated to message understanding in 1987 (Pallet 1998). Then, the Parseval/GEIG project compared phrase-structure grammars (Black 1991). A series of campaigns within Tipster project were realized on tasks like summarization, translation and searching (Hirschman 1998). In 1994, in Germany, the Morpholympics compared German taggers. Then, the Senseval and Romanseval campaigns were conducted with the objectives of semantic disambiguation. In 1996, the Sparkle campaign compared syntactic parsers in four different languages (English, French, German and Italian). In France, the Grace project compared a set of 21 taggers for French in 1997 (Adda 1999). In 2004, during the Technolangue/Easy project, 13 parsers for French were compared. Large-scale evaluation of dependency parsers were performed in the context of the CoNLL shared tasks in 2006 and 2007. In Italy, the EVALITA campaign was conducted in 2007 and 2009 to compare various NLP and speech tools for Italian; the 2011 campaign is in full progress - EVALITA web site. In France, within the ANR-Passage project (end of 2007), 10 parsers for French were compared - passage web site.

3.5.3 Different types of evaluation

Depending on the evaluation procedures, a number of distinctions are traditionally made in NLP evaluation:

Intrinsic vs. extrinsic evaluation

Intrinsic evaluation considers an isolated NLP system and characterizes its performance mainly with respect to a gold standard result, pre-defined by the evaluators. Extrinsic evaluation, also called evaluation in use considers the NLP system in a more complex setting, either as an embedded system or serving a precise function for a human user. The extrinsic performance of the system is then characterized in terms of its utility with respect to the overall task of the complex system or the human user. For example, consider a syntactic parser that is based on the output of some new part of speech (POS) tagger. An intrinsic evaluation would run the POS tagger on some labelled data, and compare the system output of the POS tagger to the gold standard (correct) output. An extrinsic evaluation would run the parser with some other POS tagger, and then with the new POS tagger, and compare the parsing accuracy.

Black-box vs. glass-box evaluation

Black-box evaluation requires one to run an NLP system on a given data set and to measure a number of parameters related to the quality of the process (speed, reliability, resource consumption) and, most importantly, to the quality of the result (e.g. the accuracy of data annotation or the fidelity of a translation). Glass-box evaluation looks at the design of the system, the algorithms that are implemented, the linguistic resources it uses (e.g. vocabulary size), etc. Given the complexity of NLP problems, it is often difficult to predict performance only on the basis of glass-box evaluation, but this type of evaluation is more informative with respect to error analysis or future developments of a system.

Automatic vs. manual evaluation

In many cases, automatic procedures can be defined to evaluate an NLP system by comparing its output with the gold standard (or desired) one. Although the cost of producing the gold standard can be quite high, automatic evaluation can be repeated as often as needed without much additional costs (on the same input data). However, for many NLP problems, the definition of a gold standard is a complex task, and can prove impossible when inter-annotator agreement is insufficient. Manual evaluation is performed by human judges, which are instructed to estimate the quality of a system, or most often of a sample of its output, based on a number of criteria. Although, thanks to their linguistic competence, human judges can be considered as the reference for a number of language processing tasks, there is also considerable variation across their ratings. This is why automatic evaluation is sometimes referred to as objective evaluation, while the human kind appears to be more subjective.

3.5.4 Shared tasks (Campaigns)

BioCreative
Message Understanding Conference
Technolangue/Easy
Text Retrieval Conference
Evaluation exercises on Semantic Evaluation (SemEval)

4.0 CONCLUSION

Systems based on machine-learning algorithms have many advantages over hand-produced rules.

5.0 SUMMARY

In this unit, you learnt:

NLP using machine learning
History of natural language processing
Major tasks in NLP
Statistical Natural Language Processing
Evaluation of natural language processing

6.0 TUTOR- MARKED ASSIGNMENT

1. List four major tasks in NLP.
2. Describe the history of natural language processing.
3. Mention different types of evaluation of NPL.

7.0 REFERENCES/FURTHER READING

Bates, M. (1995). Models of Natural Language Understanding. Proceedings of the National Academy of Sciences of the United States of America, Vol. 92, No. 22 (Oct. 24, 1995), pp. 9977– 9982.

Christopher, D. Manning, Hinrich Schütze (1999). Foundations of Statistical Natural Language Processing, MIT Press, ISBN 978-0-262-13360-9, p. xxxi.

Yucong, D. & Christophe C. (2011). Formalizing Semantic of Natural Language through Conceptualization from Existence. International Journal of Innovation, Management and Technology (2011) 2 (1), pp. 37-42.