ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Issue 11, November 2016

Computational Linguistics involved in natural language generation and understanding

Shreya Udhani

B. E, Dept. of Information Technology, Institute of Engineering and Technology, DAVV, Indore, India

ABSTRACT: Language is an ability to communicate idea. Language is an integral part to exchange ideas and views. If we are able to find a computational method of language, it will build a very robust tool for communication. We use various methods to exploit knowledge and combine these with linguistic facts to come up with a computational natural language system. There are various shortcomings of a language which we encounter but these shortcomings are actually what make the language powerful too. Language can be in form of speech or written form. The processing may therefore be in two ways. Processing the speech will be a little difficult as various challenges have to be overcome like the presence of noise or remove ambiguity. Processing of written form is relatively easier. The written text goes through lexical, syntactic and semantic analysis to extract information. This processing of written language is termed as written language processing. Natural language processing includes understanding and generation as well as translation in various languages. This makes it a crucial factor for understanding.

KEYWORDS: Discourse Integration, Morphological, Natural Language Processing, Pragmatic Analysis, Syntax, Semantics.

I. INTRODUCTION

Communication can be in any form. It can be generally written or through speech. For a two-way communication it is vital that both the objects involved have a common knowledge about the language being used. The language is then converted into knowledge and processes by us. Processing speech may be difficult than processing written form of language. This happens because there may be myriad factors that have to be taken in account to process this language. Therefore while processing written language there is a need of additional information to handle the ambiguities that arise in language. Written language processing is called Natural language processing. It is easier as it takes in account the lexical, syntactic and semantic knowledge of the language. While processing this language we come across various difficulties, but these difficulties are flip sides of the language that actually make this very robust and powerful. The difficulties that arise are-

- *Challenge*- The language gives partial description of information.
  For example, some guys are playing.

  ↓

  Some guys are playing tennis.
  2 guys are playing tennis.
  John and Tim are playing tennis.

  Positive side- Language allows us to be precise and vague. We can express only the information we require.

- *Challenge*- The language does not define the context giving rise to ambiguity.
  For example, I am reading a paper. (a research paper)
  I am reading a paper. (a news paper)

  Positive Side- Infinite information can be conveyed using finite symbols.

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Issue 11, November 2016

- *Challenge*- The language can never be complete as new words are continuously defined.
  For example, The tumor has spread to the frontal lobe.

  Positive Side- The language can be evolved and jargons can be added according to our wishes.

Language uses these challenges and converts the weaknesses into strength.
Natural language processing helps in generating and understanding the language. It is also used to translate one language into another. Understanding a language is the process of mapping the input into a more usable form of data that is to convert raw facts into information that may be used for enhancing knowledge. To understand a sentence we map it into some representation depending upon a situation. But due to wide variety of situations there is no single definition of understanding that can rightly fit in the situation. To build up a computer program that could process natural language, we have to first define the underlying task and the target representation.

II. CHALLENGES

Though it looks an easy domain to map sentences for understanding the meaning but this is not entirely true. There are various challenges that are faced during processing this. The major challenge is ambiguity of data. In English, when we communicate something, the sentence may not actually mean that thing. For example consider a sentence, Ryan's face turned blue due to the new change. This doesn't mean that his face turned blue in color, this is an expression to show that he was angry. Or another kind of ambiguity may arise when we talk about the words that have several meanings. Like, He was advised to keep his tone in check while talking to the HOD. Check means limitation or it is used in chess too. One more kind of ambiguity may come up due to affixes. For example, Jerry had many friends. It was his friend's birthday bash. The first friend refers to plural noun and the second refers to denote a third person. These challenges need a robust systematic system for processing. All these problems need to be addressed so that computer can rightly process the natural language.

III. IMPLEMENTATION STEPS

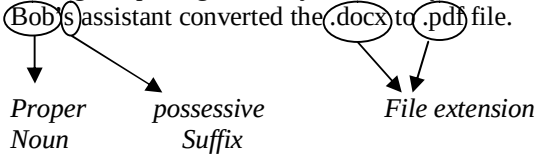
The natural language understanding process goes through five phases. These phases are:

- Morphological analysis
- Syntactic analysis
- Semantic analysis
- Discourse integration
- Pragmatic analysis

These phases have a boundary that isn't entirely clear. They may sometimes take place in a proper sequence, or sometimes all at once. When they take place in a sequence, one process may appeal for assistance to another. This can be better understood if we understand what these terms are.

i. MORPHOLOGICAL ANALYSIS

During morphological analysis, individual words are analyzed. Non-word tokens such as punctuation are separated from the words. It pulls the words apart and assigns categories. For instance,
Bob's assistant converted the .docx to .pdf file.
During morphological analysis, word by word the sentence is analyzed.



ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

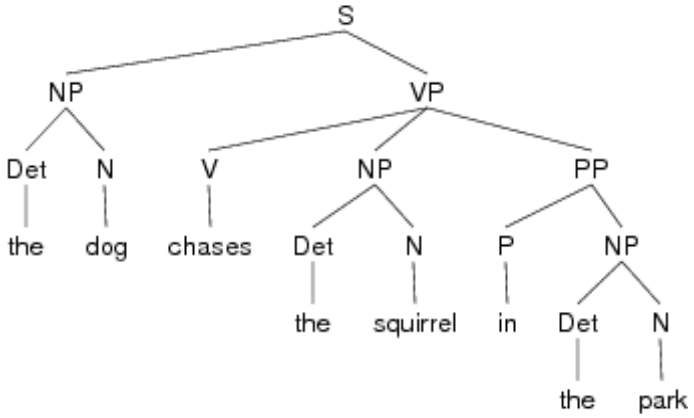
(An ISO 3297: 2007 Certified Organization)

Vol. 5, Issue 11, November 2016

As shown above the sentence is analyzed piece by piece. Every word is assigned a syntactic category. Even the file extensions are recognized in the sequence which is acting as an adjective in the above situation. In the above example the possessive suffix is recognized. This is an important step as interpretation of prefixes and suffixes will depend on syntactic category for the word. For instance, prints and print's is different. One make it plural, other makes it third person singular verb. If the prefix or suffix is wrongly interpreted the meaning of the sentence is changed. The interpretation assigns a category to the word hence removing the ambiguity from the word.

ii. SYNTACTIC ANALYSIS

Every language follows certain rules. Violation of the rules gives an error known as syntax error. The sentence is converted into structures that show correlation between the words. This correlation may violate the rules sometimes. Syntax refers to the rules the formal language has to follow. For example, "Boy the go to the store." will have a syntax error. Syntactic Analysis uses the results from morphological analysis to build the description of the sentence. The sentences divided into categories (during morphological process) are then put in defined structures. This is called parsing. For example, the dog chases the squirrel in the park, would be represented as:



Here the sentence is broken down according to categories. This is then represented in hierarchical structure with nodes as sentence units. These parse trees are parsed while syntax analysis and if an error occurs the processing stops and it shows syntax error. One more thing to keep in mind is that we can add reference markers to the entities here. This will be useful later as it will provide a space to accumulate information about the entities. The parsing may be top-down or bottom-up.

- Top-down: Begin with the start symbol and parse the sentence in accordance with the grammar rules until all the terminals of the sentence have been parsed.
- Bottom-up: Begin with the sentence to be parsed and apply rules backward until a start symbol is reached.

Sometimes these may be combined to give a hybrid parsing method.

iii. SEMANTIC ANALYSIS

The semantics deal with the meanings. It assigns meaning to every structure created by syntactic analyzer. It then maps every syntactic structure and the objects in the task domain. If no mapping is possible the structures are rejected. For example, "hot ice-cream" will produce a semantic error. While semantic analysis two major operations are performed:

- First, it will map individual words to appropriate objects in database: The entities are mapped with the database. The dictionary meaning of the entities is found. An entity can have more than one definition.
- Secondly, all these objects are integrated to find proper correlation between the structures. The process to find the correct meaning is called lexical disambiguation. It is done by associating each word with the context.

The reference markers we defined above can be used to find a partial meaning to a sentence. Semantic and syntax are two entirely disparate concepts. It may be highly likely that a syntactically correct sentence is semantically incorrect.

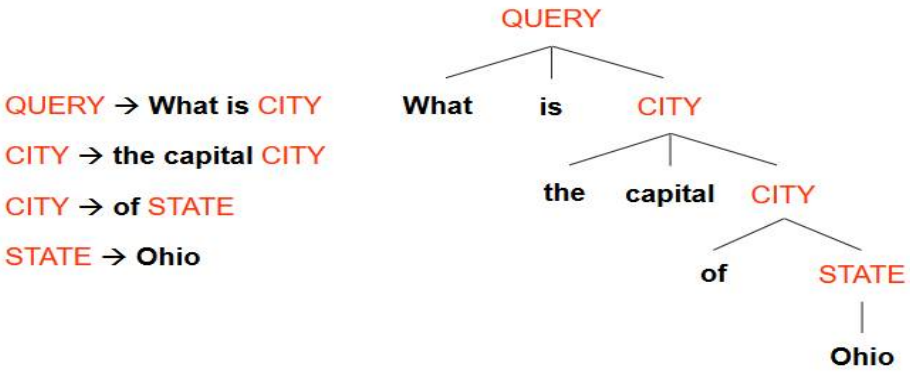
ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Issue 11, November 2016

For example, "John eats the sun." is syntactically correct as it follows all the rules of English, but it is semantically incorrect. The semantic analysis makes sure that a sentence is abiding by the rules and produces some information. It removes sentences which may seem correct but can't be used to acquire any useful information from them.



The above example shows Semantic Parsing.

iv. DISCOURSE INTEGRATION

One of the major ambiguities that arise while processing is referential ambiguity. Referential ambiguity occurs when a reference to a word can't be decided. For example, Sid ate a banana.

Tim went for a drive.

He liked it.

In the above sentence, "He" may be Sid or Tim. This creates an ambiguity. This dependency is disclosure integration. *Disclosure Integration* is defined when the meaning on an individual sentence depends upon the sentence the sentence that precedes it or follows it. Like in the above example the 3rd sentence depends upon the sentence preceding it. The model attempts to remove the referential ambiguity.

v. PRAGMATIC ANALYSIS

Pragmatic is defined as dealing with things in a more practical way rather than using theoretical approach. This rightly defines why pragmatic analysis is done. As we have discussed before a sentence may mean different things at different situations. For example, The score is 30.

↓

The score is 30. (score may be in an examination)
The score is 30. (score may be in a game)
The score is 30. (score may be in an experiment)

We see that one input have various interpretations in various situations. To understand which meaning we are talking about, we need to understand the situation. To remove this ambiguity pragmatic analysis is used.

The description of the sentence is complete after following the five analyses processes. Now, to finally understand what was said another step is taken. The structure representing is reinterpreted to determine what was actually said when this sentence was used. Like if the situation is describing an examination then first interpretation is right. For some sentences this analysis might not be needed as their intent is lucid and easily comprehensible. This analysis is required when the intent is vague. Pragmatic analysis attempts to make the intent more clear by applying a set of rules. This

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

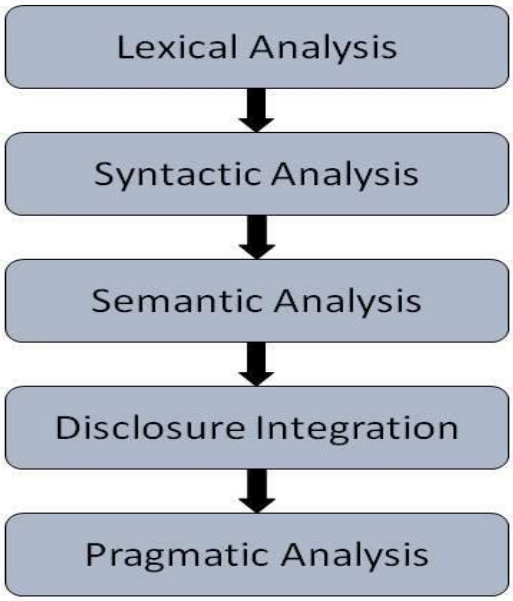
(An ISO 3297: 2007 Certified Organization)

Vol. 5, Issue 11, November 2016

produces the final meaning of the sentence. Sometimes, the final step in pragmatic analysis is to translate from knowledge based representation into an executable command.

IV. IMPLEMENTATION

The above phases or processes defined for processing are performed in sequence. Every phase takes in the input from the previous phase and sends it further for processing. An input might be rejected mid-way if it does not conform to the rules defining it.



But this might not always be the case. Sometimes the processing takes place all-together. The boundaries between these phases are vague. There might arise a situation where two or more phases are acting together. For instance, consider the sentence

Is the glass jar peanut butter?

At the end of the sentence there are four noun phrases which will be needed to form noun phrases to give a sentence of the form: “Is the x y?” where x and y are the noun phrases we need. During syntax analysis there will be following choices available:

X	Y
Glass	Jar peanut butter
Glass jar	Peanut butter
Glass jar peanut	Butter

During the syntactic analysis all of these choices look viable, but to find the correct phrases we need to analyze the semantics. When we do a semantic analysis the only choices making sense are “glass jar” and “peanut butter”. Hence, we see that these processes are separated but they can interact in various ways.

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Issue 11, November 2016

V. CONCLUSION

Language is a system which conforms to various rules. Natural language processing processes the written form of language in accordance with those rules. It attempts to remove ambiguity and make communication easier. It is used for understanding and generation of information. Natural language processing can be also used for multi-lingual translation. A sentence clause is when input in a system; it goes through various analysis procedures to be validated. A sentence if starting with a start state is then passed through various phases and finally reaches the terminal of the sentence, it is validated. A sentence is first converted into various tokens; these tokens are then analyzed syntactically by using parsing. After being syntactically correct, it is made sure that the sentence is meaningful i.e. it is semantically correct. The last phase of discourse integration and pragmatic analysis strives to find additional information that validates the sentence. After being validated, the sentence is then used to understand the information and use it in a usable form. In some cases, also translate it.

REFERENCES

[1] R. K. Ando and T. Zhang. A framework for learning predictive structures from multiple tasks and unlabeled data. Journal of Machine Learning Research (JMLR), 6:1817–1953, 2005.
[2] R. M. Bell, Y. Koren, and C. Volinsky. The BellKor solution to the Netflix Prize Technical report, AT&T Labs, 2007. http://www.research.att.com/
[3] Y. Bengio and R. Ducharme. A neural probabilistic language model. In Advances in Neural Information Processing Systems (NIPS 13), 2001.
[4] Y. Bengio, P. Lamblin, D. Popovici, and H. Larochelle. Greedy layer-wise training of deep networks. In Advances in Neural Information Processing Systems (NIPS 19), 2007.
[5] Y. Bengio, J. Louradour, R. Collobert, and J. Weston. Curriculum learning. In International Conference on Machine Learning (ICML), 2009.
[6] L. Bottou. Stochastic gradient learning in neural networks. In Proceedings of Neuro-Nımes. EC2,1991.
[7] L. Bottou. Online algorithms and stochastic approximations. In David Saad, editor, Online Learning and Neural Networks. Cambridge University Press, Cambridge, UK, 1998.
[8] L. Bottou and P. Gallinari. A framework for the cooperation of learning algorithms. In Advances in Neural Information Processing Systems (NIPS 3). 1991.
[9] L. Bottou, Y. LeCun, and Yoshua Bengio. Global training of document processing systems using graph transformer networks. In Conference on Computer Vision and Pattern Recognition (CVPR), pages 489–493, 1997.
[10] J. S. Bridle. Probabilistic interpretation of feedforward classification network outputs, with relationships to statistical pattern recognition. In F. Fogelman Soulie and J. Herault, editors, Neuro-computing: Algorithms, Architectures and Applications, pages 227–236. NATO ASI Series, 1990.
[11] P. F. Brown, P. V. deSouza, R. L. Mercer, V. J. D. Pietra, and J C. Lai. Class-based n-gram models of natural language. Computational Linguistics, 18(4):467–479, 1992a.
[12] P. F. Brown, V. J. Della Pietra, R. L. Mercer, S. A. Della Pietra, and J. C Lai. An estimate of an upper bound for the entropy of english. Computational Linguistics, 18(1):31–41, 1992b.
[13] C. J. C. Burges, R. Ragno, and Quoc Viet Le. Learning to rank with non smooth cost functions. In Advances in Neural Information Processing Systems (NIPS 19), pages 193–200. 2007.
[14] R. Caruana. Multitask Learning. Machine Learning, 28(1):41–75, 1997.
[15] O. Chapelle, B. Schlkopf, and A. Zien. Semi-Supervised Learning. Adaptive computation and machine learning. MIT Press, Cambridge, Mass., USA, September 2006.
[16] E. Charniak. A maximum-entropy-inspired parser. In Conference of the North American Chapter of the Association for Computational Linguistics & Human Language Technologies (NAACL-HLT)
[17] H. L. Chieu. Named entity recognition with a maximum entropy approach. In Conference on Natural Language Learning (CoNLL), pages 160–163, 2003.
[18] N. Chomsky. Three models for the description of language. IRE Transactions on Information Theory, 2(3):113–124, September 1956.
[19] S. Clemenc ̧on and N. Vayatis. Ranking the best instances. Journal of Machine Learning Research (JMLR), 8:2671–2699, 2007.
[20] W. W. Cohen, R. E. Schapire, and Y. Singer. Learning to order things. Journal of Artificial Intelligence Research (JAIR), 10:243–270, 1998.
[21] T. Cohn and P. Blunsom. Semantic role labelling with tree conditional random fields. In Conference on Computational Natural Language (CoNLL), 2005