OPEN ACCESS

International
Journal
Of Modern Engineering Research (IJMER)

Natural Language Ambiguity and its Effect on Machine Learning

Harish Chandra Maurya[1], Pooja Gupta[2], Nalin Choudhary[3]
[1.] *M.Tech (C.S) (Assistant Professor, Bhagwant University, Ajmer, India)*
[2.] *M.Tech (C.S) (Scholar, Bhagwant University, Ajmer, India)*
[3.] *M.Tech (C.S) (Scholar, Bhagwant University, Ajmer, India)*

ABSTRACT- *"Natural language processing" here refers to the use and ability of systems to process sentences in a natural language such as English, rather than in a specialized artificial computer language such as C++. The systems of real interest here are digital computers of the type we think of as personal computers and mainframes. Of course humans can process natural languages, but for us the question is whether digital computers can or ever will process natural languages. We have tried to explore in depth and break down the types of ambiguities persistent throughout the natural languages and provide an answer to the question "How it affects the machine translation process and thereby machine learning as whole?" .*

Keywords- *Natural language processing, ambiguity in natural language, Syntactic Ambiguities, Attachment Ambiguity, Ambiguity resolution.*

I. Introduction

Probably the single most challenging problem in computer science is to develop computers that can understand natural languages. So far, the complete solution to this problem has proved elusive, although a great deal of progress has been made. Fourth-generation languages are the programming languages closest to natural languages.

Obviously, probably it would be easier to get a computer to accomplish a task if you could talk to it in normal English sentences rather than having to learn a special language only a computer and other programmers can understand. I say "probably" because programming languages typically require you to state things in a way that is more precise than occurs in typical English sentences, and the average person might be hard-pressed to get the computer to do some mathematically precise task if he or she had to state the task in conversational English. But on the face of it, at least, it would seem to be a great thing if we could converse with computers as we do with one another. In this paper we have tried to unfold the mystery that surrounds the ambiguous nature of natural languages and their effects on machine learning and translation process shining some light on methods to overcome it.

II. Nlp: Natural Language Processing

Natural language processing (NLP) is a field of computer science, artificial intelligence, and linguistics concerned with the interactions between computers and human (natural) languages. As such, NLP is related to the area of human–computer interaction. Many challenges in NLP involve natural language understanding, that is, enabling computers to derive meaning from human or natural language input, and others involve natural language generation. Modern NLP algorithms are based on machine learning, especially statistical 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.

III. Ambiguity In Natural Language

Something is ambiguous when it can be understood in two or more possible senses or ways. If the ambiguity is in a single word it is called lexical ambiguity and in a sentence or clause, structural ambiguity.

Examples of lexical ambiguity are everywhere. In fact, almost any word has more than one meaning. "Note" = "A musical tone" or "A short written record." "Lie" = "Statement that you know it is not true" or "present tense of lay: to be or put yourself in a flat position." Also we can take the word "ambiguity" itself. It can mean an indecision as to what you mean, an intention to mean several things, a probability that one or other or both of two things has been meant, and the fact that a statement has several meanings. Ambiguity tends to increase with frequency of usage.

Some examples of structural ambiguity: "John enjoys painting his models nude." Who is nude? "Visiting relatives can be so boring." Who is doing the visiting? "Mary had a little lamb." With mint sauce?

In normal speech, ambiguity can sometimes be understood as something witty or deceitful. Harry Rusche proposes that ambiguity should be extended to any verbal nuance, which gives room to alternative reactions to the same linguistic element.

Polysemy (or polysemia) is a compound noun for a basic linguistic feature. The name comes from Greek poly (many) and semy (to do with meaning, as in semantics). Polysemy is also called radiation or multiplication. This happens when a word acquires a wider range of meanings. For example, "paper" comes from Greek papyrus. Originally it referred to writing material made from the papyrus reeds of the Nile, later to other writing materials, and now it refers to things such as government documents, scientific reports, family archives or newspapers.

There is a category, called "complementary polysemy" wherein a single verb has multiple senses, which are related to one another in some predictable way. An example is "bake," which can be interpreted as a change-of-state verb or as a creation verb in different circumstances. "John baked the potato." (change-of-state) "John baked a cake." (creation)

So far the research shows that ambiguity has something to do with the evolution of the language itself. Over thousands of years the language has evolved from a set of symbols governed by a set of rules to a much more complex phenomenon. With so many possible cases of its ambiguous nature it has become virtually impossible to figure out whether this is to our advantage or not. To get a broader idea we should look into the types of ambiguities persistent in natural languages to get a meaningful answer to this question.

IV. Types of Ambiguity

There are different types of ambiguities:

2.1 Lexical Ambiguity

It is the ambiguity of a single word. A word can be ambiguous with respect to its syntactic class. e.g.: book, study.

For e.g.: The word silver can be used as a noun, an adjective, or a verb.

- She bagged two silver medals.
- She made a silver speech.
- His worries had silvered his hair.

Lexical ambiguity can be resolved by Lexical category disambiguation i.e., parts-of-speech tagging. As many words may belong to more than one lexical category part-of-speech tagging is the process of assigning a part-of-speech or lexical category such as a noun, verb, pronoun, preposition, adverb, adjective etc. to each word in a sentence.

2.1.1 Lexical Semantic Ambiguity

The type of lexical ambiguity, which occurs when a single word is associated with multiple senses. E.g.: bank, pen, fast, bat, cricket etc.

For e.g.:

- The tank was full of water.
- I saw a military tank.

The occurrence of tank in both sentences corresponds to the syntactic category noun, but their meanings are different. Lexical Semantic ambiguity resolved using word sense disambiguation (WSD) techniques, where WSD aims at automatically assigning the meaning of the word in the context in a computational manner.

2.2 Syntactic Ambiguity

The structural ambiguities were syntactic ambiguities. Structural ambiguity is of two kinds: Scope Ambiguity and Attachment Ambiguity.

2.2.1 Scope Ambiguity

Scope ambiguity involves operators and quantifiers.

Consider the example: Old men and women were taken to safe locations.

The scope of the adjective (i.e., the amount of text it qualifies) is ambiguous. That is, whether the structure (old men and women) or ((old men) and women)? The scope of quantifiers is often not clear and creates ambiguity. Every man loves a woman. The interpretations can be, for every man there is a woman and also it can be there is one particular woman who is loved by every man.

2.2.2 Attachment Ambiguity

A sentence has attachment ambiguity if a constituent fits more than one position in a parse tree. Attachment ambiguity arises from uncertainty of attaching a phrase or clause to a part of a sentence.

Consider the example: The man saw the girl with the telescope.

It is ambiguous whether the man saw a girl carrying a telescope, or he saw her through his telescope. The meaning is dependent on whether the preposition 'with' is attached to the girl or the man. Consider the example: Buy books for children Preposition Phrase 'for children' can be either adverbial and attach to the verb buy or adjectival and attach to the object noun books.

2.3 Semantic Ambiguity

This occurs when the meaning of the words themselves can be misinterpreted. Even after the syntax and the meanings of the individual words have been resolved, there are two ways of reading the sentence.

Consider the example: Sita loves her mother and Priya does too.

The interpretations can be Priya loves Sita's mother or Priya likes her own mother. Semantic ambiguities born from the fact that generally a computer is not in a position to distinguishing what is logical from what is not.

Consider the example: The car hit the pole while it was moving.

The interpretations can be The car, while moving, hit the pole and The car hit the pole while the pole was moving. The first interpretation is preferred to the second one because we have a model of the world that helps us to distinguish what is logical (or possible) from what is not. To supply to a computer a model of the world is not so easy. Consider the example: We saw his duck. Duck can refer to the person's bird or to a motion he made. Semantic ambiguity happens when a sentence contains an ambiguous word or phrase.

2.4 Discourse Ambiguity

Discourse level processing needs a shared world or shared knowledge and the interpretation is carried out using this context. Anaphoric ambiguity comes under discourse level.

2.4.1 Anaphoric Ambiguity

Anaphors are the entities that have been previously introduced into the discourse. Consider the example, the horse ran up the hill. It was very steep. It soon got tired.

The anaphoric reference of 'it' in the two situations cause ambiguity. Steep applies to surface hence 'it' can be hill. Tired applies to animate object hence 'it' can be horse.

2.5 Pragmatic Ambiguity

Pragmatic ambiguity refers to a situation where the context of a phrase gives it multiple interpretations, one of the hardest tasks in NLP. The problem involves processing user intention, sentiment, belief world, and modals etc. all of which are highly complex tasks.

Consider the example: Tourist (checking out of the hotel): "Waiter, go upstairs to my room and see if my sandals are there; do not be late; I have to catch the train in 15 minutes." Waiter (running upstairs and coming back panting): "Yes sir, they are there."

Clearly, the waiter is falling short of the expectation of the tourist, since he does not understand the pragmatics of the situation. Pragmatic ambiguity arises when the statement is not specific, and the context does not provide the information needed to clarify the statement. Information is missing, and must be inferred.

Consider the example: "I love you too."

This can be interpreted as

- I love you (just like you love me)
- I love you (just like someone else does)
- I love you (and I love someone else)
- I love you (as well as liking you)

V. Ambiguity and Computational Linguistics

Computational linguistics has two aims: To enable computers to be used as aids in analyzing and processing natural language, and to understand, by analogy with computers, more about how people process natural language.

One of the most significant problems in processing natural language is the problem of ambiguity. Most ambiguities escape our notice because we are very good at resolving them using context and our knowledge of the world. But computer systems do not have this knowledge, and consequently do not do a good job of making use of the context.

The problem of ambiguity arises wherever computers try to cope with human language, as when a computer on the Internet retrieves information about alternative meanings of the search terms, meanings that we had no interest in. In machine translation, for a computer it is almost impossible to distinguish between the different meanings of an English word that may be expressed by very different words in the target language. Therefore all attempts to use computers alone to process human language have been frustrated by the computer's limited ability to deal with polysemy. Efforts to solve the problem of ambiguity have focused on two potential solutions: knowledge-based, and statistical systems. In the knowledge-based approach, the system developers must encode a great deal of knowledge about the world and develop procedures to use it in determining the sense of the text.

In the statistical approach, a large corpus of annotated data is required. The system developers then write procedures that compute the most likely resolutions of the ambiguities, given the words or word classes and other easily determined conditions. The reality is that there no operational computer system capable of determining the intended meanings of words in discourse exists today. Nevertheless, solving the polysemy problem is so important that all efforts will continue. I believe that when we achieve this goal, we will be close to attaining the holy grail of computer science, artificial intelligence. In the meanwhile, there is a lot more to teach computers about contexts and especially linguistic contexts.

VI. Natural Language Ambiguity and Machine Translation

The whole idea behind the machine learning process boils down to one thing, "how machines are able to translate the natural language into meaningful set of instructions with utmost accuracy and rigidity." The ambiguities mentioned above makes it extremely difficult for a machine to understand and process human language. The typical machine translation process can be defined as " the task of automatically converting one natural language into another, preserving the meaning of the input text, and producing fluent text in the output language." While machine translation is one of the oldest subfields of artificial intelligence research, the recent shift towards large-scale empirical techniques has led to very significant improvements in translation quality.

Here is a flowchart of a typical machine translation process:

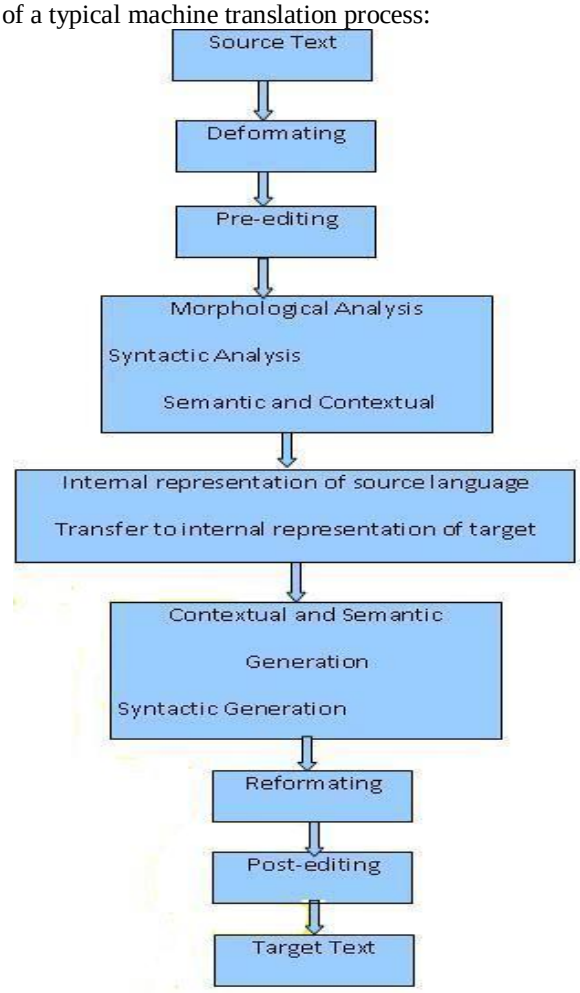

Figure 1: Machine Translation Process

6.1 Text Input

This is the first phase in the **machine translation process** and is the first module in any MT system. The sentence categories can be classified based on the degree of difficulty of translation. Sentences that have relations, expectations, assumptions, and conditions make the MT system understand very difficult. Speaker's intentions and mental status expressed in the sentences require discourse analysis for interpretation. This is due to the inter-relationship among adjacent sentences. World knowledge and commonsense knowledge could be required for interpreting some sentences.

6.2 Deformatting & Reformatting

This is to make the machine translation process easier and qualitative. The source language text may contain figures, flowcharts, etc. that do not require any translation. So only translation portions should be identified. Once the text is translated the target text is to be reformatted after post-editing. Reformatting is to see that the target text also contains the non-translation portion.

6.3 Pre-editing & Post-editing

The level of pre-editing and post-editing depend on the efficiency of the particular MT system. For some systems segmenting the long sentences into short sentences may be required. Fixing up punctuation marks and blocking material that does not require translation are also done during pre-editing. Post editing is done to make sure that the quality of the translation is up to the mark. Post-editing is unavoidable especially for translation of crucial information such as one for health. Post-editing should continue till the MT systems reach the human-like.

6.4 Analysis, Transfer & Generation

Morphological analysis determines the word form such as inflections, tense, number, part of speech, etc. Syntactic analysis determines whether the word is subject or object. Semantic and contextual analysis determines a proper interpretation of a sentence from the results produced by the syntactic analysis. Syntactic and semantic analysis is often executed simultaneously and produces syntactic tree structure and semantic network respectively. This results in internal structure of a sentence. The sentence generation phase is just reverse of the process of analysis.

6.5 Morphological Analysis & Generation

Computational morphology deals with recognition, analysis and generation of words. Some of the morphological process is inflection, derivation, affixes and combining forms. Inflection is the most regular and productive morphological process across languages. Inflection alters the form of the word in number, gender, mood, tense, aspect, person, and case. Morphological analyzer gives information concerning morphological properties of the words it analyses.

6.6 Syntactic Analysis & Generation

As words are the foundation of speech and language processing, syntax can considered as the skeleton. Syntactic analysis concerns with how words are grouped into classes called parts-of-speech, how they group their neighbors into phrases, and the way in which words depends on other words in a sentence.

The ambiguity present with the natural language processing makes it extremely difficult to get a similar and perfectly relating input and output. Usually this machine translation cycle results in 60-70 percent accuracy.

VII. Conclusion

Language cannot exist without ambiguity; which has represented both a curse and a blessing through the ages. Since there is no one "truth" and no absolutes, we can only rely on relative truths arising from groups of people who, within their particular cultural systems, attempt to answer their own questions and meet their needs for survival. Language is a very complex phenomenon. Meanings that can be taken for granted are in fact only the tip of a huge iceberg. Psychological, social and cultural events provide a moving ground on which those meanings take root and expand their branches. Signification is always "spilling over," as John Lye says, "especially in texts which are designed to release signifying power, as texts which we call 'literature'." The overlapping meanings emerge from the tropes, ways of saying something by always saying something else. In this sense, ambiguity in literature has a very dark side, when important documents are interpreted in different ways, resulting in persecution, oppression, and death.

Giving meaning to human behavior is one of the challenges for Psychoanalysis and Psychology in general. After Ferdinand de Saussure proposed that there is no mutual correspondence between a word and a thing, to ascribe significance becomes much more complicated. The meaning in each situation appears as an effect of the underlying structure of signs. These signs themselves do not have a fixed significance; the significance exists only in the individual. "Sign is only what it represents for someone." The sign appears as pure reference, as a simple trace, says Peirce. "Disambiguation" is a key concept in Computational Linguistics. The paradox of how we tolerate semantic ambiguity and yet we seem to thrive on it, is a major question for this discipline.

Computational Linguists created "Word Sense Disambiguation" with the objective of processing the different meanings of a word and selecting the meaning appropriate to the use of the word in a particular context. Over 40 years of research has not solved this problem. At this time, there is no computer capable of storing enough knowledge to process what human knowledge has accumulated. It can be seen, therefore, that ambiguity in language is both a blessing and a curse.

However after carefully studying and classifying the types of ambiguities and how it affects the machine learning process we can aim at providing solutions to improve the accuracy of the output we get in the translation cycle. Here are a few insights from our observations:

- A better human-computer interface that could convert from a natural language into a computer language and vice versa. A natural language system could be the interface to a database system, such as for a travel agent to use in making reservations. Blind people could use a natural language system (with speech recognition) to interact with computers, and Steven Hawking uses one to generate speech from his typed text.
- A translation program that could translate from one human language to another (English to French, for example). Even if programs that translate between human languages are not perfect, they would still be useful in that they could do the rudimentary translation first, with their work checks and corrected by a human translator. This cuts down on the time for the translation.
- Programs that could check for grammar and writing techniques in a word processing document.
- A computer that could read a human language could read whole books to stock its database with data.

REFERENCES

[1]. Harmain, H.M., Gaizauskas, R.: CM-Builder: A Natural Language-Based CASE Tool for Object-Oriented Analysis. Automated Software Engineering 10(2), 157–181 (2003)

[2]. Giordani, A., Moschitti, A.: Semantic Mapping Between Natural Language Questions and SQL Queries Via Syntactic Pairing. In: Horacek, H., Métais, E., Muñoz, R., Wolska, M. (eds.) NLDB 2009. LNCS, vol. 5723, pp. 207–221. Springer, Heidelberg (2010)

[3]. Mich, L., Franch, M., Inverardi, P.N.: Market research for requirements analysis using linguistic tools. Requir. Eng., 40–56 (2004) Resolving Syntactic Ambiguities in Natural Language Specification of Constraints 187

[4]. Bajwa, I.S., Bordbar, B., Lee, M.G.: OCL Constraints Generation from NL Text. In: IEEE International EDOC Conference 2010, Vitoria, Brazil, pp. 201–214 (2010)

[5]. OMG: Object Constraint Language (OCL) Standard v. 2.0, Object Management Group (2006), http://www.omg.org/spec/OCL/2.0/

[6]. Marneffe, M.C., Bill, M., Manning, C.D.: Generating Typed Dependency Parses from Phrase Structure Parses. In: LREC 2006 (2006).

[7]. http://www.coli.uni-saarland.de/projects/milca/courses/comsem/html/node92.html .

[8]. Chantree, F., Nuseibeh, B., De Roeck, A., and Willis, A. 2006. Identifying Nocuous Ambiguities in Natural Language Requirements. In Proceedings of 14th IEEE International Requirements Engineering Conference (RE'06), Minneapolis, USA, 59-68.

[9]. Berry, D. M., Kamsties, E., and Krieger, M. M. 2003. From contract drafting to software specification: Linguistic sources of ambiguity.

[10]. Dan W. Patterson, Introduction to Artificial Intelligence and Expert System , PHI, 2001, Chapter 12.

[11]. Eugene Charniak and Drew Mcdermott, Introduction to Artificial Intelligence , pearson, 1998, Chapter 4.