CSCM10 Project Report: Vague Language And Its Relevance to Natural Language Processing

Saturday 26th September, 2020

Jason Summers
SN: 903702

Abstract—The complexities of natural language can restrict the quality of interaction with computers and consequently sabotage efforts to effectively complete computer operations using speech. This report looks into the area of vague language and its implications in human computer interactions. From this a model for the types of vague language is built and its possible implementation is discussed. It is concluded that vague language holds an important role in what makes human speech human, allowing semantic warping so we may condense details for a more effective interaction.

1 Introduction

Human-computer interaction (HCI) has always been, for the most part, a somewhat one way experience with humans giving commands or queries and the computers giving basic feedback or responses. This is especially true of the foundations of computer science with programs functioning by interpreting lines of code written in a programming language. These languages usually rely on a small number of keywords and a very specific structure in order to be properly understood. This is indicative of how we interact with computers in our day to day lives. A better interaction with computers would result from the impression that the computer or program understands the context and general language of our speech in the same way humans can. Furthermore the computer would understand where we would like the interaction to go, i.e. the conversational flow. And so the uprising of Intelligent Personal Assistants (IPAs) (e.g. Amazon Alexa and Google Home) begins, these devices are designed to dialogue with owners to help perform actions such as searching the web and interacting with other services. The interest in this area of technology is increasing with the market for IPA technology being predicted to exceed $9 billion by 2023 [1].

Due to the industry relevance of this topic the parent project of this report will explore the logistics behind a more conversational interaction, primarily how machine learning and linguistic analysis can breakdown both the humans input and desired conversational flow during an inquiry. This is also known as Natural Language Processing (NLP). The idea of a more conversational interaction with an agent (an autonomous computer program designed to perform actions for a person) has the potential to be a more successful interaction [2], and to improve the human experience both through the agents success and its natural language use. In order for the conversation to seem more seamless and to fully imitate human conversations, the program will have to do its best to understand the subtle nuances in human speech. As a result of this, the topic of this report will be to understand some the subtleties behind conversational language and its use in Human-agent interaction (HAI), primarily the concept of vague language (VL).

VL is a central feature of daily language in use, both spoken and written. It can be used to express something that has a meaning that arises from intrinsic uncertainty, this is because it relies on common idiomatic knowledge and a mutual understanding of the context of the conversation [3]. By using such common phrasing with a humans intuition we are able to encode a bigger amount of feelings, expressions or knowledge into shorter sentences, and thus facilitate a smoother conversation. As a result, vagueness is not only a more natural way for us to speak but it can be a more efficient way too.

In this report we will explore a number of research articles about VL and its effect in NLP and briefly review their content. This will give us a useful understanding of the research area and its progress such that we may be up to date when entering the full topic of the parent project.

2 The meaning and importance of VL in NLP

2.1 What is vague language?

This topic plainly has its roots in the field of linguistics so we shall begin by exploring the area of VL within linguistics before asking how a program will decode these properties.

2.1.1 Some vague expressions in English, Dr J.M. Channell, 1983

Let us start with a thesis that will help us get a grasp on the fundamentals of VL as well as providing some examples. Channell delivers a deep look into the properties of VL in her 1983 thesis and speaks to its importance saying "a complete theory of language must have vagueness as an integral component" [4].

Channell states that there are many ways to be vague, one might employ "hedges" such as *virtually* or *it seems that*, or they may use agentless passives [4]. These phrases are seemingly a product of informality in that they expose the lack of required precision when simply conversing with a friend. In a formal environment, such as work, it is expected that one does not compromise a project or effort with imprecision resulting from VL. She goes on to list two distinct groups of common vagueness that lacks percision;

Type A: Number Approximations:

- *about n*
- *approximately n*
- *around n*
- *n or so*
- *n or m*

where n and m are real numbers.

Type B: 'Tag' Approximations:

- *X and things like that*
- *X and that*
- *X or anything/something like that*

where X (to be referred to as a tag for this report) may be replaced with various verbs, nouns and descriptors such as, for example, gaming, fish and smart.

From these examples we can see that vagueness is useful when delivering information involving a possible range of values or tags, something that may be resultant from genuine ignorance or reference to similar tags. No matter what the cause is, the intent is clear, the speaker wishes to give the gist rather than an exact meaning. Channell expands on this suggesting "that a vague utterance is one which cannot be assigned an exact meaning, even with recourse to context" [4]. If it is not meaning that VL asserts in the aforementioned cases then perhaps we need to look beyond the area of semantics and instead to another branch of linguistics. Channell suggests that if a universal model for understanding VL exist it must include a look into the pragmatics of the phrase as well as the semantics. This will be a key hurdle in handling VL in NLP, as analysing context will require a deeper understanding of the content of the phrase, as well as the appreciation of what has been mentioned before.

2.1.2 Pragmatic Competence: The Case of Hedging, B. Fraser, 2010

This section will look at a more recent investigation of VL, particularly the pragmatic aspect behind vagueness. Fraser focuses on the difficulties of using and understanding the pragmatics behind VL, or rather, the ability of pragmatic competence. Fraser defines pragmatic competence as "the ability to communicate your intended message with all its nuances in any socio-cultural context and to interpret the message of your interlocutor as it was intended." [5]. Pragmatic competence presents a key distinction between simply understanding the rules of a language and understanding the intended message of a phrase, and gives an indication of what makes natural language so powerful. Natural language is not fully dependant on logic and grammar but also on numerous factors such as context, tone and physical expressions or gestures. These factors can be easily implemented such that a sentence means the complete opposite to its literal meaning, something a program should acknowledge when trying to imitate conversation. Tone and physical expression will obviously be beyond what a simple text-based NLP program will be able to analyse, so what traits of language also contribute to this effect?

Fraser helps to define the area of VL by investigating the area of hedging, these areas are actually the same but we shall still evaluate Frasers outlook and compare it to our current model of VL. He gives the definition:

> "Hedging, a rhetorical strategy that attenuates either the full semantic value of a particular expression, as in *hes* ***sort of*** *nice*, or the full force of a speech act, as in *I must ask you to stop doing that*". [5]

This definition paints a more detailed picture on where and how VL affects a phrase. First of all a new variable arises from the idea of semantic attenuation, by how much is the meaning warped? and is it positive or negative attenuation? Fraser deals with this by introducing three categories to classify possible hedging:

- Propositional hedging
- Reinforcement
- Speech act hedging

Propositional hedging is the category that we have looked at the most in this report thus far. It has positive semantic attenuation, that meaning that the semantic value is weakened by hedging. Reinforcement is the opposite, it has negative attenuation, the semantic value of the phrase is strengthened. This warping of meaning is actually to the benefit of the semantic value, hence reinforcement is considered not to be a true category of hedging. Nevertheless we can still incorporate it into our model as it is still a valuable trait of natural language. The final category here is speech act hedging, this is a case where hedging is used to show the sincerity of a speech act despite the speech acts actual semantic meaning. An example of speech act hedging is "**can you** please leave?", despite being phrased like a question it is used to convey a request. Speech act hedging is an interesting area of hedging as it is used to deliver a well defined pragmatic meaning of some importance. Propositional hedging on the other hand is used in scenarios where only the gist is required.

These categories will now be the centre of our VL model as they cover informal approximations and, most importantly, the idea of a hidden meaning (semantically speaking). Reinforcement will also be a useful aspect to consider, as it has the capacity to act as a counter to some speech act hedging by reinforcing the semantic meaning of the phrase.

2.2 VL Implementation

Now that we have a good understanding of the concept of VL, and the ways it can enhance natural language, let us now look at how it concerns computer processing. In this section we will first breakdown some of the fundamentals behind typical NLP, then we will see how VL can be implemented. It should be noted however that the focus of this report is not to fully investigate the state of modern NLP but merely to understand the basics.

2.2.1 Natural Language Processing (Almost) from Scratch, R. Collobert et al, 2011

We shall be looking at two categories of methods in NLP that we can use, those which describe syntactic information (e.g., part-of-speech tagging, labelling and parsing) or semantic information (e.g., word-sense disambiguation, semantic role labelling, named entity extraction and anaphora resolution) [6]. Collobert et al test each method to determine their effectiveness, we shall briefly look at each method and determine the ideal method to implement our model for VL. Part-of-speech tagging (POS) aims to label each word with a unique tag based on its syntactic role, e.g. plural noun, adverb, etc [6]. There is room for additional procedures to help raise the per-word accuracy for tag identification, the use of maximum entropy classifiers and inference in a bidirectional dependency diagram has produced accuracies as high as 97.24% [6][7].

Chunking works in a similar fashion but instead it first labels segments of sentences with syntactic constituents and then assigns each word a tag depending on if it begins the chunk or not. Chunking has shown slightly worse results than POS but only by a few percent (F1 scoring)[6].

Methods based on semantic information such as named entity recognition and semantic role labelling not only show notably worse F1 scores [6] but are also at the most risk from VL, for this reason we will focus on the syntactic information methods in this report.

Given the grouping nature of chunking, it may be an ideal candidate to choose for our NLP method, one which would allow classification of VL phrases. This may be done by introducing another dimension to each tag, this dimension could specify how the semantic value of the tag changes, based on other chunks and tags from previous phrases. The model could be trained on example phrases of hedging such that it can recognise what common words and syntax change the semantic value of the sentence. For instance, in the phrase "there are 5 *odd*", an example of propositional hedging, the word *odd* can be given an extra label indicating that it likely contributes to VL. Finally we could adapt the model, most likely using a cost function, such that it can identify the semantic value for the whole phrase.

Speech act hedging may be a little more difficult, an initial point that we can start from would be to evaluate how literally the program should take a speech act. Employing a similar system to the propositional hedging could also work for word such as *guess*, *must* and *should* as these words are usually not taken as being literal in cases of VL.

The actual implication of VL in code will have to be investigated in the main CSCM10 project but the methods mention thus far should help to provide a good starting point.

2.2.2 Exploring VL use and voice variation in HAI, Dr Leigh Clark, 2016

In this section we look at the significance of implementing VL in NLP, particularly how vagueness can increase the ability for an agent to converse with a human and ultimately seem more human-like. Clark outlines many of the key ideas of this report with the addition of a new VL tool, referred to by Trappes-Lomax as a minimiser [8]. These minimisers can be divided into three separate forms, *like, basically* and *just* [9], and serve to simplify a speech act. This may have one of two affects depending on context. Either they reduce the assertiveness of the statement or request by downplaying the importance of its content, or they increase the aggression by expressing how simple a request is. The last effect is more situational and has links to tone as well as context so for now we shall ensure we are aware of this and move on.

Clark investigated the effect of incorporating VL into HAI by having participants build Lego models under the guidance of a model agent using vague and non vague language, the participants were then asked various questions to determine the quality of their experience. The results outlined the importance of something not explored in this report but seemingly just as important as VL, this being a human voice. When asked if a human voice would be preferred 75% of the participants in the vague agent group said yes with only 42% in the non vague group. Whilst there were some concerns with the quality of the text-to-speech recordings, many participants made remarks on the overly robotic nature and insincerity of the voice.

It seems that in order to fully utilize VL for agents to imitate human speech, one must also ensure that the agents voice also imitates human speech to the best of its ability (one might even take this further by giving the agent a face).

3 Conclusion

This report has explored the meaning of VL and its potential use in NLP and ultimately HAI. There is much more to be said about HAI however, for VL is but one aspect of what makes human conversation so fluid and natural. Nevertheless it seems that there is potential for its implementation if we can ensure that the rest of the agent interface also imitates humans.

Due to the vastness of the area of natural language this report focussed on developing a model for the composition of VL, outlining two key groups, propositional and speech act hedging. These core groups, along with some other notable tools such as reinforcement and minimisers provide a strong foundation for NLP. This report has room for improvement, as seen, VL is potentially useless or even distracting without implementing realistic tonal variations, this reasoning is also applicable to this report. Furthermore testing of the ideas put forward here are lacking, empirical validation would strengthen the ideas put forward.

References

[1] "Global intelligent virtual assistant market 2018-2023:," https://www.businesswire.com/news/home/20180723005506/en/Global-Intelligent-Virtual-Assistant-Market-2018-2023-Market.

[2] N. M. Radziwill and M. C. Benton, "Evaluating quality of chatbots and intelligent conversational agents," *arXiv preprint arXiv:1704.04579*, 2017.

[3] J. Cutting, *Vague language explored*. Springer, 2007.

[4] J. M. Channell, "Vague language: Some vague expressions in english," Ph.D. dissertation, University of York, 1983.

[5] B. Fraser, "Pragmatic competence: The case of hedging," *New approaches to hedging*, vol. 1534, 2010.

[6] R. Collobert, J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu, and P. Kuksa, "Natural language processing (almost) from scratch," *Journal of machine learning research*, vol. 12, no. Aug, pp. 2493–2537, 2011.

[7] K. Toutanova, D. Klein, C. D. Manning, and Y. Singer, "Feature-rich part-of-speech tagging with a cyclic dependency network," in *Proceedings of the 2003 conference of the North American chapter of the association for computational linguistics on human language technology-volume 1*. Association for Computational Linguistics, 2003, pp. 173–180.

[8] H. Trappes-Lomax, "Vague language as a means of self-protective avoidance: Tension management in conference talks," in *Vague language explored*. Springer, 2007, pp. 117–137.

[9] L. M. Clark, "Exploring vague language use and voice variation in human-agent interaction." Ph.D. dissertation, University of Nottingham, 2016.