Towards Identifying Ontological Semantic Defaults with Big Data: Preliminary Results

Tatiana Ringenberg, Julia Taylor, John Springer
Computer & Information Technology
Purdue University
West Lafayette, IN, USA
{tringenb, jtaylor1, ja}@purdue.edu

Victor Raskin
Linguistics & CERIAS
Purdue University
West Lafayette, IN, USA
vraskin@purdue.edu

***Abstract*—This paper reports on a work-in-progress and suggests a method of detecting conceptual defaults in natural language big data. It combines Hadoop and Nutch technologies for web crawling with the Ontological Semantic Technology (OST) in an initial effort of this kind. Initial results demonstrate the viability of this method to detect unintended inference within text.**

Keywords-big data, Hadoop, Nutch, conceptual default, Ontological Semantic Technology.

I. INTRODUCTION

This paper merges big data research technology with the important computational semantic task of identifying conceptual defaults, i.e., the parts of text that the speaker/writer omits because they are too obvious to mention, both for himself/herself and for their intended audience—and occasionally for all. Thus, hardly anybody would say, *I unlocked the door with the key*, preferring to drop the prepositional phrase as obvious [1]-[3]. The prepositional phrase is, then, a conceptual default. On the other hand, if a competent speaker does verbalize a default, the hearers may suppose that the prepositional phrase is not the default: for instance, if all the locks in the building are electronic. The defaults are very important for the computer to be aware of and use for inferences and reasoning as people do.

Though little to no work has been done on defaults outside of Ontological Semantics Technology (OST), defaults can be loosely related to Grice's Maxim of Quantity which states that a person will not mention more than what is necessary in a conversation [11]. This definition is quite broad. The work in this research seeks to solve the problem of identifying a small portion of information which a speaker considers to be too trivial to mention.

The algorithm used in this research is intended to pull only unintentional inference related to verb events, nouns and adjectival modifiers. In previous work, this specific type of default has been referred to as White Dude Inference (WD-Inference) [2].

Big Data is used in this research as a method of confirmation of our implementation of the WD-Inference algorithm. The website purdue.edu was chosen due to the abundance of texts. The website also fits very well with the principles of Big Data including high volume, velocity and variety of information assets [13]. As such the data is prevalent, and varied, enough to test the algorithm.

This paper describes the implementation decisions and steps in implementing the algorithm for basic WD-Inference-style defaults. Section II describes the background information related to Ontological Semantics, Ontological Semantics Technology and Big Data. Section III describes the materials used for implementation of the algorithm. Section IV describes the methods used in the preliminary study. Section V describes preliminary results of the research. Section VI describes the next steps that will be taken to improve and complete the research. Section VII describes future work related to Ontological Semantics defaults and OST.

II. BACKGROUND

A. Conceptual Defaults

Surprisingly little has been done about conceptual defaults before [1]-[3]. These papers could not have emerged without the Ontological Semantic Technology [4]-[9] that emerged from Ontological Semantics of the 1990s [10], and made it possible to develop the comprehensive human-like meaning text representation of text on the basis of an engineered language-independent ontology and a set of language-specific lexicons, whose every sense of every entry is anchored in ontological concepts linked with multiple properties. A pattern-matching and a graph-producing analyzer, the central POST elements, compete with each other in Text Meaning Representation (TMR) production. Figure 1 shows the OST architecture.

It would be a stretch to consider Grice's [11] Maxim of Quantity or scarce work on semantic ellipsis [12] as bearing on defaults but it may provide some comfort to a novice because these sources do bear somewhat on what is not necessary to say and how the text that is there may help to reconstruct the text that was elided, in some cases, because it was obvious.

To the knowledge of the researchers, no computational implementation of defaults has previously been created either. As such, there is no significant work to which we may compare the performance or the accuracy of the proposed solution. The solution provided in this research is intended to be a first step towards automatic acquisition of Ontological Semantics defaults.

 ISBN: 978-1-61208-390-2

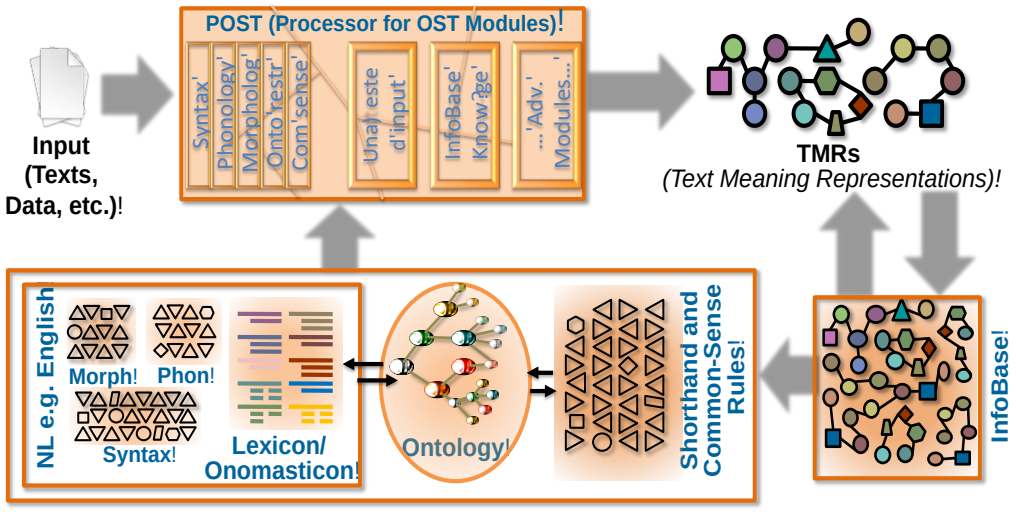

Figure 1. OST Architecture

B. Big Data

According to Gartner, Inc. [13], "Big data in general is defined as high volume, velocity and variety information assets that demand cost-effective, innovative forms of information processing for enhanced insight and decision making." As depicted in Figure 2, at the heart of Big Data is analysis and refinement leading to more effective decision-making.

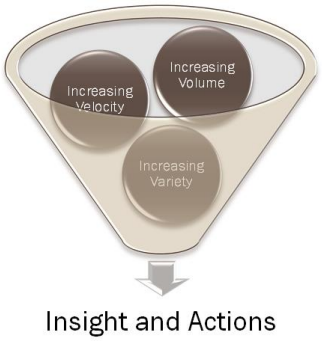

Figure 2. Big Data's 3V Leading to Insight and Actions

Moreover, Big Data lies at the confluence of several fields and disciplines – including Computation and Cyberinfrastructure, Visual Analytics and Visualization, Ethics, and Quality Assurance (see Figure 3) – and in its most effective application, has a context in a particular domain such as Business/Finance, Social Sciences, Life Sciences, Physical Sciences, and Engineering. It is in leveraging the intersection of all of these areas that Big Data delivers its greatest value.

Big Data is also pervasive. According to a report from McKinsey [14], "[l]eaders in every sector will have to grapple with the implications of big data, not just a few data-oriented managers. The increasing volume and detail of information captured by enterprises, the rise of multimedia, social media, and the Internet of Things will fuel exponential growth in data for the foreseeable future."

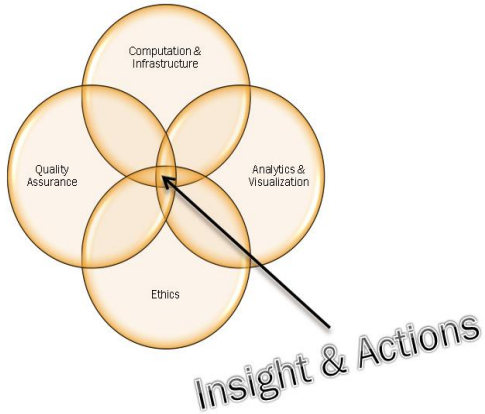

Figure 3. Confluence of Fields and Disciplines

Big Data projects cover a gamut of areas and uses. These include such well-known scientific projects as the Large Hadron Collider and the Large Synoptic Survey Telescope as well as the daily operations of Facebook and Google.

Given its dimensions related to variety and velocity, Big Data has an obvious relationship with Natural Language Processing as natural language represents a frequently generated source of "unstructured" data.

To capitalize on this ideal source of Big Data, we leverage a technology synonymous with Big Data, Hadoop, and one of the solutions for web crawling, Nutch, built on it. By leveraging these tools, we lay the foundation for scalability beyond small data sets for our NLP needs.

C. Purpose

The overall goal of this research is to provide a mechanism for the identification of a very small subset of

defaults. Specifically, this research examines the relationships between events (verbs) and the nouns and adjectives related to them. It is hypothesized that defaults should not show up within text unless modified. This means that the default for *drive* is *car* (the entity that is being driven), *car* should very rarely show up in text with the verb *drive* by itself. The only time it would be acceptable to see *car* with *drive* would be when the speaker does not generally drive a car or when a description of a car is provided. For instance, it is unusual to say *I drive a car*, because that's too trivial, but *I drive a red car* is fine, especially, if this information about a car is new.

In this study, we examine the 200 most common verbs in Brown corpus along with the noun and adjective arguments that accompany them. We compare the list of verbs with no arguments, verbs with a noun argument and verbs with a noun and adjective argument in order to find and analyze potential defaults. Future work will examine other types of defaults.

III. SELECTION OF MATERIALS

A. Parser Tool Selection

In order to pull verbs, adjectives and nouns from verb phrases a parser was required. The goal for these parsers was to allow the researcher to pull adjectives and nouns that modified a particular verb. Stanford Parser 3.4.1 was chosen for its popularity within Computational Linguistics and its parsing flexibility [16].

B. Corpus and Verb Selection

As this research seeks to analyze the relationships between verbs, nouns and adjectives specifically, a tagged corpus was ideal for the preliminary stages of research.

Brown Corpus was specifically chosen for this task because of its size, part-of-speech tagging and wide use within Computational Linguistics. Brown Corpus is a collection of American English documents from the 1960's. It consists of about 500 samples and around a million words [15].

Using Brown Corpus, all verbs were pulled from the sentences and stemmed using Porter Stemmer. Stemmed verbs that the researchers believed would not provide relevant information were removed from the list of verbs. The verbs that were removed include *say, be, go, get, "have", state*, and their forms. Verb frequencies were then calculated and the 200 most frequent verb stems were selected.

C. Structure Selection

Stanford Parser was used to generate typed dependencies. Typed dependencies are used to find relationships between words in a sentence. The goal is to provide syntactically and (partial) semantically useful information about the relationships between words. The following is a dependency for the sentence *Jackie Brandt singled deep into the hole at the short to start the rally*:

nn(Brandt-2, Jackie-1)
nsubj(singled-3, Brandt-2)
nsubj(start-11, Brandt-2)
root(ROOT-0, singled-3)
advmod(singled-3, deep-4)
det(hole-7, the-6)
prep_into(singled-3, hole-7)
prep_at(hole-7, short-9)
aux(start-11, to-10)
xcomp(singled-3, start-11)
det(rally-13, the-12)
dobj(start-11, rally-13)

Dependencies were chosen as the primary tool for sentence analysis due to simplicity. Though Dependency Grammars are not as expressive as syntax trees, they make the relationships between verbs, adjectives and nouns more transparent.

D. Crawl Selection

Brown Corpus consists of documents created in the 1960's. As such, it is the researchers' belief that a more up-to-date corpus is needed to confirm the relevance and accuracy of the methodology described in this paper.

It is also our belief that this methodology must be scalable. Given the large number of blogging and social networking venues, data these days is both prevalent and large. As such, we apply our methodology to a larger dataset than Brown Corpus.

In order to confirm the methodology for extracting defaults, described in this paper, Hadoop and Nutch were used. As was mentioned earlier, Hadoop is often used to work with Big Data and Nutch provides a very easy and intuitive web crawling experience that goes well with Hadoop.

The website "purdue.edu" was used as the initial seed for the crawl. The Purdue website was chosen due to its terms-of-service, the vast amount of data that is associated with a university and the variety of textual content. University websites, especially starting at the main site, consist of large html and text documents. This was perfect for this analysis as this research is only focused on text.

In configuring Nutch for crawling, no prefixes were excluded from the crawl. However, only html and text documents were pulled. A depth of 10 was used to limit the number of links the crawl would follow. A maximum number of 6,000 sites was used to limit the size of the data.

IV. PRELIMINARY METHODS FOR PULLING DEFAULTS

A. Scope

This research seeks to identify semantic defaults for verb events only. Any events that are not nouns are outside of the scope of this research and will be addressed in future work.

B. Procedures

1. Identify the most frequent 200 verbs from Brown corpus.

2. Pull Sentences using the Verb List. Once the top 200 verb stems are chosen, all of the sentences containing those verbs in any verb forms (for example, VBG, VBZ, VBP) are taken from Brown Corpus. This means that if a sentence had

 ISBN: 978-1-61208-390-2

the word *walking* tagged as a verb, the sentence was pulled. If the sentence had the word *walking* tagged as a noun, the sentence was not pulled. This resulted in 14719 sentences.

3. Create Dependency Representations. Using Stanford Parser, dependency representations were created for each sentence.

4. Select Noun and Adjective Arguments for Verb Events. For this initial analysis, we chose to select the following:

- all lone verbs;
- verbs with just nouns attached to them;
- verbs with adjectives and nouns attached.

As this was the only information needed from the dependency grammars, we chose to only select lines of the dependency grammars with the tags "nsubj", "dobj", "iobj" and "amod". The "nsubj tag was used to pull verbs, "dobj" and "iobj" were used in order to connect a verb to a noun ,or to pull verbs that did not have an "nsubj" tag, and "amod" was used to connect the verb-noun combinations to an adjective.

The dependency grammar below demonstrates how information was pulled:

nn(Brandt-2, Jackie-1)
nsubj(singled-3, Brandt-2)
nsubj(start-11, Brandt-2)
root(ROOT-0, singled-3)
advmod(singled-3, deep-4)
det(hole-7, the-6)
prep_into(singled-3, hole-7)
prep_at(hole-7, short-9)
aux(start-11, to-10)
xcomp(singled-3, start-11)
det(rally-13, the-12)
dobj(start-11, rally-13)

In this example, if *singled* were one of the top 200 verbs, it would be pulled because it has the "nsubj" tag. Originally we could classify it as a lone verb. *Start* would be pulled from the dependency grammar for the same reasons. "Start" and *rally* would also be pulled because of the "dobj" tag. However, because *start* is paired with *rally* in dobj, "start" would be removed from the lone verb list and added to the list of verbs with nouns.

5. Compare Verb-Noun List to Verb-Noun-Adjective List. In this step, we compared the list of verbs with nouns to the list of verbs with nouns and adjectives. As no default should appear unmodified, the verb-noun combinations from the verb-noun list were removed from the verb-noun-adjective list. Thus, if a noun occurred with a verb and had no modifier, it could not be considered a possible default for that event. For example, if we had the verb event *eat* we would see it with the unmodified noun *food food* is placed on the candidate list of defaults for *eat*. However, since we know that eating food is not informative, we don't expect to see the verb-noun pair (eat, food) to occur. It is entirely possible to say I eat hot food. This is because the heat of the food is relevant and not implied or assumed. Thus, we expect that it is possible to see the triple (eat, food, hot). If we saw the triple (*eat, food, hot*) in the corpus and never saw (*eat, food*) alone in the corpus, we would flag *food* as the default for *eat*. However, if we did see (*eat, food*) alone in the corpus then we would not consider *food* to be the default for *eat*.

V. Discussion of Preliminary Results

In pulling the relevant information from the dependencies, it was found that there were 13435 instances of lone verbs that map to events. The verbs with the highest frequency of lone occurrence included *made, come, felt, knew, began* and *look*. It was also found that there were 8240 verb-noun combinations and 2565 verb-noun-adjective combinations. Examples of verb-noun combinations included *reduce expense, create resources* and *wrote parts*. Examples of verb-adjective-noun combinations included *reported local romance, need new box* and feel questioning eyes. There were 2235 instances of candidate defaults found for 449 unique verb forms.

Although further analysis must be done, this seems consistent with [1-3] on defaults and the WD-Inference. According to their work, a verb-adjective-noun combination would likely represent the violation of a default. This is because the indication of additional detail in an event points to information that is out of the ordinary for the speaker or writer. As such, we would except to see the fewest instances of these combinations. We can see this in the example *feel questioning eyes*. The fact that the author needed to indicate that the eyes were questioning implies that something is out of the norm. This is consistent with how a native speaker would see that phrase. It is implied to the native speaker that *questioning eyes* are out of the ordinary.

The abundance of lone verbs is also consistent. When we say *I drove* the implication is that we drove a car. However, we don't actually mention the car unless it is out of the ordinary. For instance, if I am used to driving a motorcycle then it would be significant for me to say *I drove a car*.

VI. Next Steps

The next step in this research will be to apply the same methods used on Brown Corpus to the data that was pulled from the Purdue web crawl. The data will then need to be compared to the Brown sample to determine whether or not the findings from the structures are consistent.

In looking at the Web Crawl, we plan to pull data from not just the text documents themselves but also from image titles as well. It is possible that defaults and default violations will exist in this data.

Once both corpora have been examined, the verbs, nouns and adjectives will need to be acquired into the Ontology and Lexicon. Text Meaning Representations will then be generated for these sentences.

 ISBN: 978-1-61208-390-2

VII. FUTURE WORK

A. Implementation into OST

In order to fully implement default detection into OST, methods will need to be created for storing defaults. There will also need to be methods created for flagging defaults within a Text Meaning Representation. This may possibly require the creating of an InfoBase for each individual contributor.

B. Creation of InfoBase-like structures for individual defaults

As this is initial research concerning defaults, we are examing the defaults of a group of authors. Ideally, we need to be able to pull a set of defaults for a single author. As of now, we believe this will require individual InfoBases. InfoBases are meant to show the connections between several TMRs in order to create a larger picture of a conversation. Recording a series of defaults and default violations for an individual will help us better understand both what a person is saying and not saying in a conversation.

REFERENCES

[1] J. M. Taylor, V. Raskin, C. F. Hempelmann, and S. Attardo, "An unintentional inference and Ontological property defaults," Proc. IEEE_SMC, Istanbul, Turkey, 2010

[2] V. Raskin, J. M. Taylor, and C. F. Hempelmann, "Ontological semantic technology for detecting insider threat and social engineering," Pre-Proc. NSPW-10. Reprinted in: K. Beznosov, ed., Proceedings: New Security Paradigms Workshop 2010. September 20-23, 2019, Concord, MA, USA. New York: ACM Press, 2010

[3] V, Raskin, and J. M. Taylor, "A fresh look at semantic natural language information assurance and security: NL IAS from watermarking and downngrading to discovering unintended inferences and on to situational conceptial defaults," in: B. Akhgar and H. R. Arabnia, eds., Emerging Trends in Informationa nad Communication Technologies Security. Amsterdam: Elsebvier (Morgan Kaufmann, 2013

[4] V. Raskin, C. F. Hempelman, and J. M. Taylor, "Guessing vs. knowing: The two approaches to semantics in natural language processing,: Proc. Dialogue 2010. Bekasovo/Moscow, Russia, pp. 642-650, 2010

[5] J. M. Taylor, C. F. Hempelmann, and V. Raskin, "On an automatic acquisition toolbox for ontologies and lexicons in ontological semantics," Proc. ICAI-10, Las Vegas, NE, pp. 863-869, 2010

[6] J. M. Taylor, and V. Raskin, Understanding the unknown: Unattested inpiut in natural language," Proc. FUZZ-IEEE-11. Taipei, Taiwan 2011

[7] C. F. Hempelmann, J. M. Taylor, and V. Raskin, "Application-guided ontological engineering, " Proc. ICAI-10. Las Vegas, NE, 2010

[8] J. M. Taylor, V. Raskin, and C. F. Hempelmann, "From disambiguation failures to common-sense knowledge acquisition: A day in the life of an ontological semantic system," Proc. WI-IAT-11. Lyon, France, 2011

[9] J. M. Taylor, V. Raskin, and C. F. Hempelmann, "Towards computational guessing of unknown word-meanings: The ontological semantic approach," Proc. CogSci-11. Boston, MA 2011

[10] S. Nirenburg, and V. Raskin, Ontological Semantics. Dordrecht: D. Reidel, 2004.

[11] H. P. Grice, "Logic and conversation," in: P. Cole and J. L. Morgan, eds., Syntax and Semantics, Vol. 3, Speech Acts. New York: Academic Press, 1975

[12] M. J. McShane, A Theory of Ellipsis. New York: Oxford University Press, 2005

[13] J. Manvika, M. Chui, B. Brown, J. Bughin, R. Dobbs, C. Roxburgh, C., and A. H. Byers, Big Data: The Next Frontier for Innovation, Competition, and Productivity.

[14] Big data. *IT Glossary*, in Gartner: Retrieved November 21, 2014, from http://www.gartner.com/it-glossary/big-data, 2014

[15] W.N. Francis, and H. Kucera, (1979). Brown corpus manual. Brown University.

[16] M.C. De Marneffe, and C.D. Manning, (2008). Stanford typed dependencies manual. URL http://nlp. stanford. edu/software/dependencies manual. pdf.

 ISBN: 978-1-61208-390-2