CS447 Literature Review: Part of Speech and Semantic Role Labeling in Non-Canonical Grammars

Matthew Kennedy,
mk65@illinois.edu

November 28, 2023

Abstract

In ideal environments, we are often given well-structured canonical grammars that consist of sentences that follow a set of grammatical rules. How does using non-canonical data, which doesn't follow predefined sets of rules for natural language, impact methods of semantic role labeling and part-of-speech tagging? Additionally, how does lexical canonicalization, semantic intent analysis, and just leaving the grammar as non-canonical affect the semantic labels and parts-of-speech given to a grammar?

1 Introduction

In classroom and learning environments, we are often given "ideal" datasets, such as the famous MNIST digits dataset that most people use for their first deep-learning network, or WordNet for NLP tasks. Ideal datasets most commonly offer canonical grammar, which consists of canonical clauses that are defined by Huddleston and Pullum's "A Student's Introduction to English Grammar" as, "those which are syntactically the most basic or elementary". A more colloquial definition of canonical grammar would be a grammar that consists of standard sentence structures that follow the typical and expected rules of a language. In English, this canonical grammar is what is often taught in formal education settings and used in formal writings, such as this paper. In other languages, such as Korean, which we'll explore more later in this writing, the language has a structure that can be difficult to predict or interpret for a machine. While most of the language we see in educational and professional settings may follow these aforementioned rules of canonical grammar and sentences, what we often see online or in more relaxed, casual settings is **non-canonical grammar**. This type of grammar can be defined as consisting of sentences that deviate from the standard rules and patterns. Examples may be existential clauses (There's a man outside), or left dislocations (My wife, she doesn't know this) (Pullum), or other languages that have difficult-to-predict rules. Non-canonical grammar is what we most often see in settings such as social media; however, it can also be found in formal works, such as in the writings of Charles Dickens (Pham), or in entire dialects and languages (Götz).

When a given dataset consists of a grammar that follows standard, canonical rules, the tasks of **semantic role labeling** and **part-of-speech tagging** can utilize this structure to better assist in deciding what can be labeled. However, when these rules are not guaranteed, and sentences can follow any given variety of structure, these tasks become more difficult and involves the challenge of deciphering conversational skills and sentiment within a sentence. Throughout this paper, we will explore several different articles written on the topic of semantic role labeling and part-of-speech

tagging on non-canonical data - one might say, natural language in the wild. We will look at a few different approaches, including working in a language with limited data and different grammatical rules, breaking down complicated sentences into simpler ones, and wondering if we even *need* to worry that data isn't canonical.

2 What to do about non-canonical language?

One of the first questions that we must consider on this topic is what counts as "canonical" data, and where did this standard come from? Looking at our first article, "What to do about non-standard (or non-canonical) language in NLP" by Barbara Plank gives us this consideration immediately. Plank recounts that the Penn Treebank Wall Street Journal (WSJ) corpus in the 80s has become a standard for many tasks in the NLP field, but why? Is it because journalists from the WSJ are less susceptible to writing errors, and follow a codified standard? Plank continues by asking, if NLP had emerged more recently, such as the last 10 years, what would the de-facto standard be today? It's possible we would have an "inverted" world, where social media is a standard and other sources we consider "better" today are "bad language". The author is trying to demonstrate that our definitions of "canonical" data are historical coincidence and motivated by resources. All sources have their own biases and set of errors to deal with (Plank, p.1).

Plank continues by showcasing that processing non-canonical data is difficult across domains, with domains being a somewhat arbitrary term here to what is relevant to a given application. However, does annotation and the focus of this paper, semantic role labeling and part-of-speech tagging, also struggle across domains and with non-canonical data? Plank states there are 3 main approaches for working with non-canonical data at the time this paper was written.

The first is to just annotate more data to work with. However, this is naive for a few reasons. First, we must face the problem of training data sparsity, where the cross product of domain (whatever that may be) and language (whatever that may comprise) is largely uncovered by the available data. Additionally, the domain and language we are looking at is almost certainly a small subset of all languages and domains. Furthermore, communication changes, languages and dialects evolve, so what we annotate today compared to what we process tomorrow can vary drastically (Plank, p.2-3).

The second approach is making data resemble each other more. One way to do so is *normalizing* the data to make this data closer to what is expected by the technology; this is something that will be discussed more in section 4. Lastly, we have domain adaptation, where a model trained on some source domain is adapted to a different domain. Approaches here can vary, and often utilize feature augmentation, shared representation learning, instance weighting, and other steps, but almost share an unrealistic assumption that the target domain is known and available, and this sample of target data is used to adapt a given model. An example of adapting model domains is cross-lingual learning between multiple languages, or training data on the curated, canonical WSJ dataset then asking it to perform semantic role labeling on random tweets (Plank, p.3).

Plank proposes the concept of fortuitous data, which is data that is usable with some refining, is available and easy to use, and is largely user generated data that may not, and is most likely not, canonical by any standards. This data includes text such as Wikipedia entries, websites, social media posts, and more. Specifically with semantic role labeling, Plank states that hyperlinks could be used to build more robust named entity and part-of-speech taggers, or HTML markup for parsing, or mining Wiktionary for large pools of unambiguous data instances (Plank, p.3-4).

Variety	Sample	TnT	$BILTY_{\vec{w}}$	$BILTY_{\vec{w}+\vec{c}}$	OOV
(in-do.)	wsj.test	96.63	96.25	97.85	20
Domain	answers	90.08	91.24	91.93	27
	emails	91.03	89.81	92.20	29
	Tw (foster)	90.25	92.47	92.26	28
	Tw (oct27)	65.98	66.37	67.16	52
age	U35	86.11	85.06	96.53	20
	O45	86.73	85.81	87.70	22
language	da	35.25	37.85	38.00	89
	pt	24.99	43.50	47.33	93
	sv	33.13	39.80	37.09	92

Table 1: Tagging accuracy on various test set varieties (domains, languages and age groups; Tw=Twitter), using coarse POS (Petrov et al., 2012). OOV: out-of-vocabulary rate wrt WSJ.TRAIN. Accuracy is significantly correlated with OOV rate (p = -0.70) (Plank, p.5).

2.1 Testing

Plank uses two POS taggers - TnT[6], an HMM-based tagger, and BILTY, a bidirectional LSTM tagger, both of which were trained on the WSJ training portion converted to Universal POS tags. For testing, the following sources of data were used: emails and answers from the Web Treebank set, two Twitter datasets (FOSTER and GIMPEL/OCT27, or Twitter sample 1 and 2 respectively), review data from two age groups (above 45 and below 35), and data from the CoNLL-X dataset from other Indogermanic languages.

2.2 Results

The results in table 1 indicate that swapping domains from the canonical data (wsj.test) to non-canonical data (answers, emails, Tw (foster), and Tw (oct27)), had largely promising results. Tagging accuracy was highest for the WSJ test set, which is unsurprising since this is the canonical data the model was trained with. However, the interesting results for this paper are the next few domains - the answers, emails, and first twitter set, which consist of a variety of data sources and formats, and fall under non-canonical data compared to what these models were trained on. While the accuracy is not as high as the WSJ test set, accuracy is still promising at over 90% for these models. Additionally, we can see that the OOV percentage is slightly higher than the WSJ test set (Plank, p.5-6).

With some model variety and fortuitous data, these results could show promising improvement, and shows us that non-canonical data has promise with the task of semantic role analysis. Using concepts introduced in this paper such as domain adaptation and variety, model variance, and user-generated non canonical data, we can provide competitive analysis versus a strictly canonical trained-and-tested model.

Now that we have seen an overarching analysis of how non-canonical data can be useful, even using a model trained on canonical data, let's explore an article that tries to find intent and dive into semantics for non-canonical data

3 Machines Getting with the Program: Understanding Intent Arguments of Non-Canonical Directives

In the paper "Machines Getting with the Program: Understanding Intent Arguments of Non-Canonical Directives" by Cho et al. (2020), Cho et al. discuss how discerning the intention from a conversation vs non-canonical question or command is still a challenge. An example of this might be "you know where we should go today", which contains no punctuation and fails to follow most canonical standards of natural language. Human listeners can interpret the subject of this inquiry as "the destination of today", but this is a challenge for a machine (Cho et al., 2020, p.1-2).

Cho et al continue by discussing how designing dialog systems, which are systems intended to converse with a human, and more specifically their subcomponent dialog manager, which is responsible for the state and flow of the conversation, face a different set of challenges when interacting with non-canonical data. Additionally, Cho et al discuss more complex challenges for languages outside English, particularly ones that use a distinguished syntax or do not use Latin-like alphabets, or a language that is less explored and has less data available. Cho et al focuse on Korean, which has a morphology that is agglutinative, a syntax that is head-final, and scrambling, which is non-deterministic permutations of word/phrase ordering, which is common for native speakers. With the perspective of improving the performance of these aforementioned dialog managers, Cho et al propose that being able to extract intent argument of non-canonical directives will improve the parsing accuracy of these systems using a structured language query extraction scheme (Cho et al., 2020, p.2-4).

3.1 Dataset

To briefly summarize some of the previously mentioned challenges with extracting intent from a language like Korean, Cho et al. create a corpus of 50,000 Korean question/command intent pairs, with an emphasis on classifying the utterance type, or purpose of the statement, in order to improve the process of semantic role labeling. Korean is a language that suffers from a lack of data available, and can also vary in statement structure, which means it can be largely classified as non-canonical (Cho et al., 2020, p.4-5).

3.2 Results

Using this corpus, Cho et al. utilize two different seq2seq models (an RNN-based encoder-decoder with self-attention, and a transformer) to evaluate how well semantic intent can be extracted from non-canonical data, with intent being categorized into several specified categories, such as yes/no questions. Listed below in Table 2 are the results of both models, with the transformer utilizing the highest train-test split performing best. We can see all models performed well in semantic role labeling for each prompt, and ultimately finding the intent of prompts given for non-canonical data (Cho et al., 2020, p.6-8).

In the prior paper by Plank (section 2), we saw one author utilize non-canonical data that was already available for part-of-speech tagging. In this paper, we see Cho et al. (2020) generate non-canonical data in Korean to label semantic roles of sentences and ultimately find intent in order to better improve dialog managers and AI conversations with real people. Next, let's take a look at an article where authors attempt to simplify sentences in order to better label core verb semantic roles within a sentence.

	RNN S2S + Attention	Transformer	
Test Split	9:1	7:3	9:1
Iteration	100,000	10,000	10,000
ROUGE-1	0.5335	0.5383	0.5732
BERTScore	0.7693	0.8601	0.9724
Total	0.6514	0.6992	0.7728

Table 2: Test results, with test split, number of iterations, ROUGE-1 score, BERTScore, and Total, which is the average of the ROUGE-1 score and BERTScore. Scores listed are related to finding semantic intent from prompts (Cho et al., 2020, p.8).

4 Sentence Simplification for Semantic Role Labeling

In the articles we've reviewed so far, we have seen one set of authors work with this non-canonical (fortuitous was the term used) data at-hand (Plank), and another author generate non-canonical data in another language Cho et al. (2020). In the next article, Vickrey and Koller (2008) discusses simplifying complex sentences and breaking them into smaller, canonical parts in order to better label semantic roles in these smaller chunks of a sentence. Rather than accepting data as-is, Vickrey and Koller attempt to take complicated, non-canonical sentences, and use parse-trees to break these sentences into smaller, canonical chunks that can be used to infer semantics.

While in more recent years, natural language systems and models, and computing in general, continue to get more powerful and could work with non-canonical data even after being trained on a canonical data set (Plank), this was not always a fortunate advantage available to researchers. In 2008, most large language models struggled to properly label the semantic roles of sentences outside of the domain they were trained on. In fact, training on any kind of non-canonical data, or data that didn't follow a set of grammatical rules, was almost always a difficult task. Vickrey and Koller (2008) explores that current semantic role labeling systems at the time this paper was written relied primarily on syntactic features being parsed to classify roles. However, sentence structure may not always be similar, and finding training data that mimics test sets with the multitude of possible examples and sentences is a fruitless endeavor due to the number of possible sentences (Vickrey and Koller, 2008, p.1-2).

Instead of simply using a large, difficult to handle non-canonical sentence that likely has no accurate representation in a training set, Vickrey and Koller (2008) opt to create a mapping from full, complicated sentences to simplified sentences in order to easily label semantic roles that most likely have representation in the training set. Their method of breaking complex sentences into simple ones involves hand-written syntactic simplification rules with machine learning to determine which rules to prefer. After the sentence is simplified, this sentence is fed as input to a SRL system to label semantic roles (Vickrey and Koller, 2008, p.2).

While this approach does make for improved semantic labeling results, it also does lose some semantic value. Vickrey and Koller provide an example where the sentence "I was not given a chance to eat" is simplified via several rules to become "I ate", which are semantically very different sentences. However, Vickrey and Koller make the important distinction that while some semantic information will be lost, the authors are only interested in labeling core arguments of the verb, which in this example sentence would be "I" - the authors note that this simplification is much closer to

canonicalization over summarization, and the overall goal is not to provide a single shorter sentence given an input sentence, but rather, for each verb in the sentence to produce a "simple" sentence which is in particular canonical form relative to that verb (Vickrey and Koller, 2008, p.2-3).

4.1 Transformation Rules

The authors describe the approach of transformation rules, which takes as input a parse tree of the sentence and produces as output a different, changed parse tree, based on the rules written by the authors. The rules applied to each transformation step creates constraints at each node in the parse tree, where the transformation step modifies matching nodes in the parse tree to ultimately simplify sentences and canonicalize them. The authors use significant pieces of "machinery" in the rule set - for example, one is a *floating node*, which is used for locating an argument within a subordinate clause. Overall, there are 154 mostly unlexicalized rules for this approach, with the goal of these rules to avoid conservative ones, i.e. rules that have low precision (Vickrey and Koller, 2008, p.3-5).

4.2 Results

The authors generated a dataset based on possible sentence/target verb pairs s, v, and use this potentially large set of candidate labelings, defined as C^{sv}, to train a probabilistic model. Additionally, Vickrey and Koller define a scoring function based on which rules were used to simplify a sentence, which role pattern was used, and features about the assignment of constituents to roles, after which a log-linear model assigns probability to each simple labeling equal to the normalized exponential of the score. Applying this scoring function in conjunction with the dataset generated, the results are shown below for the model in table 3, where "Baseline" is a comparison to a strong Baseline SRL system that learns a logistic regression model using the features of Pradhan et al. (2005), the authors' model is labeled as "Transforms", and the "Combined" model is a hybrid model which combines the Baseline and Transforms models by falling back on the Baseline model if the Transforms model does not have any strong predictions. Additionally, the top-performing SRL system at the time of writing this paper is shown at the bottom of table 3 on row "Punyakanok" (Vickrey and Koller, 2008, p.5-7).

Model	Dev	Test WSJ	Test Brown	Test WSJ+Br
Baseline	74.7	76.9	64.7	75.3
Transforms	75.6	77.4	66.8	76.0
Combined	76.0	78.0	66.4	76.5
Punyakanok	77.35	79.44	67.75	77.92

Table 3: F1 Measure using Charniak parses (Vickrey and Koller, 2008, p.7)

Looking at table 3, we can see that the authors' Transforms model outperforms the Baseline model, and on several test sets, the combined model has further improvement. The author notes that their Transforms model performs particularly well on sentences with very long parse paths, such as "Big investment banks refused to step up to the plate to support the beleaguered floor traders by buying blocks of stock, traders say". Some potential future improvements the authors note include augmenting the rule set to handle more constructions and doing further sentence normalizations, or incorporating parser uncertainty into the model, where the simplification system is capable of seamlessly accepting a parse forest as input (Vickrey and Koller, 2008, p.8).

Some interesting concepts were introduced in this paper, where a model utilized sentence parse trees and sentence simplifications in order to better label semantic verb roles. For the last article of this paper, we are revisiting and expanding on the approach suggested by Plank of fortuitous data - we visit a different article by Plank, where they suggest that we simply don't need to modify non-canonical data at all.

5 Non-canonical language is not harder to annotate than canonical language

As discussed in section 2, Plank discussed how our view of "canonical" data, and what we consider "standard" for tasks in the NLP field, is usually WSJ or newswire or something similar, as opposed to non-canonical data such as twitter. Recall that Plank points out our standards are based on historical coincidence and not some journalistic code-of-writing. In the paper "Non-canonical language is not harder to annotate than canonical language" by Plank et al. (2015), Plank provides a response to a critique which a prior NLP research article received where the use of natural language data from twitter as opposed to a more canonical source like newswire could be "problematic" (Plank et al., 2015, p.1).

Plank continues this train-of-thought regarding why we consider some sources more canonical - if we consider newswire or other news sources "better" because writers are trained and are less likely to produce errors, what do we consider canonical in languages that don't write newspapers? Or is newswire canonical because it's what corpora are made of and the only data that was available to the NLP community for a long time?. Plank assures that this paper is not just a fight of words - the authors are addressing the issue that labeling text as non-canonical alludes to the potential challenge if these texts are used for NLP tasks. Most corpora, such as semantic treebanks, are human-annotated subsets of newswire due to historical availability of newswire (Plank et al., 2015, p.1-2).

Plank et al. review that non-canonical language can have *processing* challenges associated, such as more mixed language, more ad hoc spelling conventions, and texts directed at smaller audiences with more knowledge required during interpretation. However, newswire also comes with complexities, such as headlines, creative language, citations, and more. The authors pose that, if we ignore the skewed distribution of language linguistic resources due to historical use of newswire and other canonical sources, why should processing social media be harder than processing newswire? This skew underlines the need for newer resources, and raises the question of whether annotating and labeling non-canonical language is inherently harder than annotating and labeling canonical language (Plank et al., 2015, p.2).

5.1 Experiment

Plank et al. set up a sample experiment where two expert annotators reviewed five different corpora with fifty sentences each, where each corpora has different degrees of perceived canonicity. The corpora are listed below - note that authors indicate the WSJ corpora has the highest degree perceived canocity, and Twitter had the least.

1. Wall Street Journal (WSJ): Section 23 from the Ontonotes distribution of the Wall Street Journal dependency treebank (Bies et al., 2012; Petrov and McDonald, 2012).

2. Answers: The Yahoo! Answers test section from the English Web Treebank (Bies et al., 2012; Petrov and McDonald, 2012).

3. Spoken: The Switchboard corpus section of the MASC corpus (Ide et al., 2008).

4. Fiction: The literature subset of the test section of the Brown test set from CoNLL 2008 (Surdeanu et al., 2008), which encompasses the fiction, mystery, science-fiction, romance and humor categories of the Brown corpus.

5. Twitter: The test section of the Tweebank dependency treebank (Kong et al., 2014).

The experiment consists of two expert annotators who were tasked with identifying the main predicates and arguments in sentences across these five domains. Specifically, they were asked to find the main verb (MV), the subject (A0), and either the direct object if there is a MV, or the attribute in a copula construction (A1). The only guideline provided to the annotators was not to mark auxiliaries, and that the first word in a coordination or multiword unit is the head (Plank et al., 2015, p.2-3).

	Match		F1			
Domain	Exact	Frames	A0	A1	MV	Micro
WSJ	66%	82%	0.87	0.66	0.83	0.79
Twitter	52%	66%	0.91	0.69	0.79	0.80
Answers	74%	84%	0.98	0.81	0.88	0.90
Spoken	43%	74%	0.91	0.56	0.88	0.79
Fiction	64%	78%	0.83	0.75	0.79	0.80

Table 4: Agreement statistics between the two annotators. Micro is the micro-averaged F1 scores of the three tasks assigned to the annotators (Plank et al., 2015, p.3)

5.2 Results

From the results seen in table 4, Plank et al. showcase that WSJ had average rates of agreement between the two annotators, and the Answers dataset actually had the highest by far. Additionally, the micro-averaged F1 score for Answers performed best here, with the WSJ corpora performing near equally to the other corpora. Part of this could be due to sentence length and complexity; the WSJ corpora has longer, more complicated sentences, whereas Answers had less complicated and shorter sentences. However, this is a small toy experiment, so any results would need more thorough investigation. This experiment does provide insight into the potential that overall, what we perceive as canonical vs non-canonical data each have their own sets of biases we must account for when using.

6 Discussion

We have seen a variety of approaches from different authors. We reviewed Plank simply accepting non-canonical data as-is and embracing the seemingly endless amounts available to work with. Cho et al. (2020) proposed a method of creating a corpus in a different language that has more difficult

grammatical rules to follow, and training natural language systems to correctly analyze semantic intent from these statements. We visited Vickrey and Koller (2008) showcasing an approach of breaking down more complicated sentences into simpler ones, in order to better identify verb semantics. And lastly, we revisit Plank et al. (2015) who again ask if non-canonical data is truly more difficult to work with, or simply a different set of biases.

In each of these papers, we have seen promising results in the non-canonical domain. In section 2, we saw several part-of-speech taggers perform impressively well on domains the models were completely untrained on, such as Yahoo! Answers, Twitter, and different languages. For section 3, the authors created a corpus of non-canonical Korean statements, and all 3 models had impressive results, with the most-trained Transformer model performing best at identifying the intent of a statement. In section 4, which is our oldest paper coming in at 2008, we observe the authors breaking apart sentences and removing extra/unnecessary information to simplify sentences and better identify semantics. When this approach was combined with the leading-model at the time of authorship, we see model post improved results, particularly with sentences that are longer and more complicated. Lastly, we see in section 5 a proof-of-concept toy experiment, that non-canonical language can be equally difficult to annotate when compared to what we consider canonical data.

The idea of working with what could be a less curated, even less "ideal" dataset can be daunting when approaching a natural language task. However, throughout this paper, we have seen a variety of approaches working with non-canonical data and posting successful results in every task. It may seem more standard to work with newswire or WSJ or something common, but in general, we can see from these results that working with data outside of the domain we might be comfortable in, such as social media or another language, can improve models.

7 Conclusion

The driving question of this paper was, "How does using non-canonical data, which doesn't follow predefined sets of rules for natural language, impact methods of semantic role labeling and part-of-speech tagging?". In section 2, we saw that non-canonical domains of data only slightly hindered a model that had no training on these domains. In section 3, the authors display how training a model on a non-canonical dataset can perform well with semantic intent analysis. From section 4, we are shown how simplifying sentences can compete with standard models from the time the paper was written, and improve these same models when working in conjunction. Lastly, from section 5, we are shown a model experiment than non-canonical language is similarly difficult to canonical language, such as newsview, to annotate. In all of these experiments, we see non-canonical language competing with canonical metrics, and sometimes even outperforming, showing that what we might not consider standard for a corpus can actually be quite useful in natural language tasks.

References

Won Ik Cho, Youngki Moon, Sangwhan Moon, Seok Min Kim, and Nam Soo Kim. 2020. Machines getting with the program: Understanding intent arguments of non-canonical directives. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, page 329–339, Online. Association for Computational Linguistics.

Sandra Götz. Non-canonical syntax in south asian varieties of english: A corpus-based pilot study on fronting. In *Zeitschrift für Anglistik und Amerikanistik*.

Teresa Pham. Hard to beat dickens' characters": Non-canonical syntax in evaluative texts. In *Zeitschrift für Anglistik und Amerikanistik*.

Barbara Plank. What to do about non-standard (or non-canonical) language in nlp.

Barbara Plank, Héctor Martínez Alonso, and Anders Søgaard. 2015. Non-canonical language is not harder to annotate than canonical language. In *Proceedings of the 9th Linguistic Annotation Workshop*, page 148–151, Denver, Colorado, USA. Association for Computational Linguistics.

Geoffrey Pullum. Canonical sli.

David Vickrey and Daphne Koller. 2008. Sentence simplification for semantic role labeling. In *Proceedings of ACL-08: HLT*, pages 344–352, Columbus, Ohio. Association for Computational Linguistics.