<table>
<thead>
<tr>
<th><strong>Title</strong></th>
<th>Engineering an aligned gold-standard corpus of human to machine oriented Controlled Natural Language</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Author(s)</strong></td>
<td>Hazem Safwat; Brian Davis; Manel Zarrouk</td>
</tr>
<tr>
<td><strong>Publication Date</strong></td>
<td>2018-12-03</td>
</tr>
<tr>
<td><strong>Publication Information</strong></td>
<td>Safwat, Hazem, Davis, Brian, &amp; Zarrouk, Manel. (2018). Engineering an aligned gold-standard corpus of human to machine oriented Controlled Natural Language. Paper presented at the IEEE/WIC/ACM International Conference on Web Intelligence (WI2018), Santiago, Chile, 03-06 December, doi: 10.1109/WI.2018.00-58</td>
</tr>
<tr>
<td><strong>Publisher</strong></td>
<td>IEEE</td>
</tr>
<tr>
<td><strong>Link to publisher's version</strong></td>
<td><a href="https://dx.doi.org/10.1109/WI.2018.00-58">https://dx.doi.org/10.1109/WI.2018.00-58</a></td>
</tr>
<tr>
<td><strong>Item record</strong></td>
<td><a href="http://hdl.handle.net/10379/14953">http://hdl.handle.net/10379/14953</a></td>
</tr>
<tr>
<td><strong>DOI</strong></td>
<td><a href="http://dx.doi.org/10.1109/WI.2018.00-58">http://dx.doi.org/10.1109/WI.2018.00-58</a></td>
</tr>
</tbody>
</table>
Engineering an aligned gold-standard corpus of human to machine oriented Controlled Natural Language

Hazem Safwat  
Insight Centre for Data Analytics  
National University of Ireland, Galway  
hazem.abdelaal@insight-centre.org

Brian Davis  
Insight Centre for Data Analytics  
National University of Ireland, Galway  
brian.davis@insight-centre.org

Manel Zarrouk  
Insight Centre for Data Analytics  
National University of Ireland, Galway  
manel.zarrouk@insight-centre.org

Abstract—Knowledge base creation and population are an essential formal backbone for a variety of intelligent applications, decision support and expert systems and intelligent search. While the abundance of unstructured text helps in easing the knowledge acquisition gap, the ambiguous nature of language tends to impact accuracy when engaging in more complex semantic analysis. Controlled Natural Languages (CNLs) are subsets of natural language that are restricted grammatically in order to reduce or eliminate ambiguity for the purposes of machine processability, or unambiguous human communication within a domain or industry context, such as Simplified English. This type of human-oriented CNL is under-researched despite having found favor within industry over many years. We describe a novel dataset which aligns a representative sample of Simplified English Wikipedia sentences with a well known machine-oriented CNL. This linguistic resource is both human-readable and semantically machine interpretable and can benefit a variety of NLP and knowledge based applications.

Index Terms—Natural Language Processing, Controlled Natural Language, Knowledge Extraction, Semantic Web

I. INTRODUCTION

Knowledge base creation and population is of paramount importance for a variety of applications including decision support and expert systems. Formal knowledge may also act as semantic backbone for language processing and information retrieval applications. A knowledge base for a particular domain may be either non existent or incomplete. While the abundance of unstructured text offers a means of easing the knowledge acquisition gap, the highly ambiguous nature of language impacts on accuracy when moving beyond the entity level to more complex semantic interpretation of text for knowledge creation. Controlled Natural Languages (CNLs) are unambiguous natural languages based on a restricted grammar that map into formal knowledge structures [1]. CNLs are an attempt to circumvent linguistic ambiguity and have found applications with respect to formal knowledge creation, ontology development and domain specific knowledge based machine translation [2]. Aside from specific knowledge gathering activities, which by the vary nature of the task, demand a restricted natural language interface for knowledge creation by domain experts, CNLs offer little incentive to the average user to create formal knowledge, even as implicit side effect of an authoring effort involving semantic annotation or enrichment of text content. A subcategory of CNL which offers a middle ground of reduced ambiguity for semantic interpretation but less restriction than a machine-oriented CNL. Human-oriented CNLs [3] have been in wide spread use for many years. Their development was motivated by the purposes of language learning, and unambiguous communication between humans in a domain context. An example of human-oriented CNL is the Simplified text such as Simple Wikipedia3. It is a form of text written using style guides2 to reduce complexity and ambiguity of the language, especially for non-native speakers and juniors. Some of the written style guides recommended for authors in Simple Wikipedia include: use active voices, avoid compound sentences (e.g conjunctions), avoid idioms (multi-words), keep sentences short and informative. Other human-oriented CNLs include ASD Simplified Technical English3, developed to improve the readability and comprehensibility in technical documents. In addition to, Boeing Technical English to improve the communication between people for air traffic control [4]. The development and planning of these CNLs appears often community driven, many of which at first glance, like Simplified, may been inspired by Basic English5. Human-oriented CNLs unlike machine-oriented CNLs [3] do not attempt to unambiguously map into formal knowledge structures, as the communication goal is human to human and not human to machine. An interesting premise is how close linguistically such languages are to machine-oriented CNLs and if so to what degree? Moreover, rewriting all or most of a human-oriented CNLs into a machine-oriented CNL which can be unambiguously parsed into a formal knowledge structure [17] could unlock significant silos of implicit general purpose domain knowledge, contained within existing human-oriented CNL content.

This paper focuses on our initial experiments with respect to the computational linguistic analysis of a corpus of human-oriented CNL represented in Simplified English, as well as the

1https://simple.wikipedia.org/wiki/  
3http://www.asd-ste100.org/
investigation of the feasibility of rewriting Simplified English into a well known machine-oriented CNL called ACE [7]. Since no ground truth exists, it has been necessary to engineer one for our experiments.

We present a linguistic resource that is both human-readable and semantically machine interpretable. This resource is a snapshot taken from the abstracts of the Simple English Wikipedia dump\(^4\). The selected abstracts are rewritten into a machine-oriented CNL, by applying some rules on the syntactic structures of the sentences to be accepted and parsed by the CNL semantic parser. To our knowledge this resource is the first human to machine CNL aligned dataset. The paper is structured as follows: Section II describes related work, Section III presents the corpus collection, processing and analysis. In Section IV, we discuss the results and the evaluation, and finally, Section V offers a conclusion.

II. RELATED WORK

Attempto Controlled English (ACE) is a CNL developed for knowledge representation. We choose ACE for our experiments for the following reasons: 1) It is a widely adopted CNL with an expressive grammar. 2) The language can be automatically mapped into different formal languages such as Discourse Representation Structures (variant of first order logic) [7] and subsequently a subset of ACE can be converted to the Web Ontology Language (OWL). 3) It also provides access to different tools and resources to use the language such as the ACE parsing engine - APE\(^5\), which we use for our validation.

The work from [8] analysed the text in Simple Wikipedia and Wikipedia linguistic features to produce a new resource corpus that can help sentence simplification research. It provides a new simplification dataset that is an improved version over Simple Wikipedia. Their analysis over the text in both Simple Wikipedia and Wikipedia was done for the purpose of showing the amount of simplification between two parallel sentences from both Wikipedias, and introduce a new comparative approach to simplification corpus analysis. The main difference between our work is that we analyse both Wikipedias for the purpose of rewriting simplified English into a formal knowledge using CNLs as they tend to be less ambiguous.

In [9] the authors present a text rewriting approach to increase the amount of labeled data available for model training. They analysed Simple Wikipedia and Wikipedia parallel corpora to automatically extract rewrite rules, then generate multiple versions of sentences annotated with gold standard labels for the purposes of semantic role labeling. Our work is similar to them in that we also use rewrite rules to generate gold standard resource. However, the main difference is that we use rewrite rules to generate CNL sentences from Simple Wikipedia sentences, which means our efforts are directed towards a different purpose.

In [10] the author, presents a semantically annotated corpus developed using a bootstrapping approach using NLP techniques for parsing and semantic interpretation, together with an interface for collaborative annotation of experts and crowdsourcing community. Although they generated a semantic resource with deep semantics, the resource is small as it includes less than 5k sentences which is much smaller than ours.

The other resources of ProPBank [11] and Framenet [12] are semantically annotated corpora. Although, both of the resources are valuable to the community, they lack the deep semantic representation that combines different layers of semantic annotation into one formalism [10].

III. CORPUS ANALYSIS

We first collected a dataset of the abstracts from the Simple Wikipedia dump and its parallel Wikipedia dump. All experiments are performed on Simple Wikipedia abstracts only and their corresponding Wikipedia abstracts. Table I, describes the style guides from Simple Wikipedia on how to write simplified text versus several common stylistic properties observed across several CNLs (both human and machine oriented) [13]. From the table we can see that both texts significantly overlap in most of the properties.

A. Corpus Collection and Pre-processing

We collected the XML formatted dumps containing both Simple Wikipedia abstracts and their corresponding Wikipedia abstracts. Then, we converted the XML format into JSON\(^6\) format. After that we cleaned the text using regular expressions which includes removing special characters, remove incomplete or blank sentences and abstracts, text between brackets, etc. In table II, we show all the steps performed for the pre-processing of the dumps.

The total number of abstracts before cleaning was around 74k in the Simple Wikipedia dump. Since we had not yet cleaned the dump, we did not count the parallel Wikipedia abstracts. After cleaning the data, we extracted around 48.8k abstracts from Simple Wikipedia and in parallel we found around 27.5k abstracts in Wikipedia as some of the abstracts are found blank, incomplete or missing. The total number of sentences extracted from the cleaned Simple Wikipedia was around 87k, including 968.2k tokens, with most of the abstracts including two sentences. In parallel, the total number of sentences extracted from the cleaned Wikipedia dump was around 39.2k, including around 586.7k tokens. In table III, we show a comparison between the two dumps before and after cleaning.

In order to test our intuition with respect to rewriting simplified text represented in the Simple Wikipedia abstracts to a machine-oriented CNL, we decided to analyze the presence of CNLs features in both within text abstracts of Simplified Wikipedia and the corresponding Wikipedia abstracts, the purpose being to measure how similar the Simplified text

\(^4\)https://dumps.wikimedia.org/simplewiki/
\(^5\)https://github.com/Attempto/APE
\(^6\)Java Script Object Notation
to the common CNL features identified in [13]. In regard to the feasibility of rewriting Simplified Wikipedia sentences to a CNL, our first assumption is that the simplified text should be logically less ambiguous and less complex than standard Wikipedia unstructured text. Consequently, its linguistic properties will overlap significantly more with CNLs than unstructured text. So in order to measure this, we analyzed some of the measurable properties from [13] that should be present in the Simplified English text, given that it could be rewritten into CNL. As shown in table IV, the first metric is the length of sentences (number of tokens/sentence). In Simple Wikipedia we found that more than 90% of the sentences does not exceed the 20 tokens. On the other side, sentences from the parallel Wikipedia abstracts are usually exceeding this limit. Although the guidelines for writing Simple Wikipedia abstracts recommended the authors to avoid using passive voices, we found that 34% of the sentences did not follow this rule. On the other hand, 51% of the sentences in Wikipedia articles, are written using the passive voice. Gerunds are the words that are formed with verbs but act as nouns e.g go swimming, were found to be 6% in Simple Wikipedia sentences and 21% in Wikipedia sentences. CNLs usually use determiners before nouns, so our test found that the tags which preceded nouns in the Simple Wikipedia sentences are ranked as follows: 1) Determiners, 2) Noun Phrases, 3) Prepositions, 4) Adjectives, but in the Wikipedia sentences the list was different as follows 1) Nouns, 2) Noun Phrases, 3) Prepositions, 4) Determiners. Moreover, noun clusters are found in 4% of the Simple Wikipedia sentences and 8% in the Wikipedia sentences. Hence, based on the observations above, we can confirm our hypothesis that the linguistic properties of Simplified Wikipedia text overlap more with CNLs than unstructured text.

Based on the results above, we conducted a deeper analysis of the Simple Wikipedia POS structures which overlapped completely with the CNL rules. This meant extracting all abstracts which follow the common CNLs rules in table IV from the original dataset dump, excluding the remainder. As shown in table V, the total number of abstracts from the Simple Wikipedia dump that follow the CNL rules in Table IV are found to be around 20.4k. These abstracts include around 36.5k sentences, with 383.5k tokens. The result is our
IV. RESULTS

Since Simplified English rules and style guides request authors to use preferred sentence forms such as Subject-Verb-DirectObject, and Subject-Verb-IndirectObject\(^7\), our second hypothesis is that most of the sentences, which have the same POS tags structure/pattern (forms), can be rewritten into CNL using a few rules without changing their semantics. So, we created a dictionary that includes the POS tags structures of all the sentences in the corpus. The main aim of creating this dictionary is to discover to which extent the authors of the Simple Wikipedia abstracts followed the guidelines of writing preferred sentence structures for writing the simplified text. So, we grouped all similar POS tags structures together and count them, in order to estimate the percentage of unique POS tags structures and the percentage of repeated POS structures in the corpus. This would thus help us approximate the number of rewrite rules needed to map Simplified Wikipedia sentences into CNL. We found total number of 22,083 unique POS tag structures. Next, we estimated the number of sentences that belongs to a group of POS tags structures. As shown from table VI, we present 5 cases from the dictionary. For example, the dictionary shows that 102 POS tags structures include around 7.8k sentences, and 629 POS tags structures include around 12.6k sentences. This analysis indicated that there are a lot of repeated POS structures that could be found in the corpus.

We extracted the top 5 repeated POS tags structures in the corpus and we show our analysis on them in table VII. We found that the first structure is found 634 times in the corpus. We experimented with rewriting a few sentences which match this POS structure into ACE. So, in order to rewrite this POS tags structure, we need to chunk the noun clusters into one noun, to be accepted by the APE parser and validated as an ACE sentence for automatic translation into DRS formalisms. Although POS tags structures number 2, 3 and 5 are not as frequent as POS tags structure number 1, they did not require any rewriting to be accepted by the ACE parser or converted into DRS as they tend to be very basic sentences. The POS tags structure number 4, can rewritten using a noun phrase chunker in the beginning of the sentence to chunk the names as one noun, and another noun phrase chunker in the end of the sentence to chunk the combination consists of a noun after an adjective into one noun. Although table VI shows that there are some repeated POS tags structures in the corpus, still the number of sentences to POS tags structures ratio (12,630 sentence /630 POS tags structure) is not high, if we compared it to the total size of the corpus which is around 36.5k sentences. This means that in order to rewrite 12,630 sentences into CNL we need to implement rules that can cover the 630 different POS tags structures, which is a lot of rules that will lead to rewriting only 35% of the corpus. Although it was shown from table VII that some few basic sentences did not require a rule for rewriting these sentences, the table showed that the percentage of these is still very low.

A. Annotation and Validation

So, in order to overcome this problem we found from our analysis that a common rule from table VII contains noun clusters and/or adjective noun clusters - noun phrase chunks effectively. So, we applied chunking rules on all the sentences in the corpus and assigned them a new POS tag COMP. A new set of new POS tags structures are created and the results from this analysis is summarized in table VIII. In the table we notice that around 19.2k sentences are captured together under 300 POS tags structures only, out of total 13,867 POS tags structures. This means that if we can rewrite the 300 structures into ACE CNL, we envisage having around 19.2k sentences in ACE CNL format and consequently into the DRS formal representation and exported to OWL.

we applied a total of 10 rules\(^8\) to the sentences, where some of the results of these rules can be shown in table IX. The table shows the top 5 POS tags structures after the chunking took place. The highest occurring POS tags structure was found 2664 times. The rewriting rule for this structure is to chunk the nouns at the end of the sentence. The second most dominating structure was found 1399 times, and the rewriting rule chunked the nouns at the beginning and of the sentence. Structure number 3 occurred 1088 times, and the rewriting rule chunked the nouns at the end of the sentence. Structure number 4 occurred 926 times and the rewriting rule chunked a noun/group of nouns, followed by POS tag IN, followed by a noun/group of nouns. The last structure occurred 885 times, where the past verb of the sentence is rewritten into the present tense and the nouns at the beginning and end of the sentence are chunked.

\(^8\)https://drive.google.com/open?id=1eBUZXzH7iEJHJ3EptqG4acii4–

B. Evaluation

The evaluation of the generated resource is performed on two sides, 1) The system coverage, and 2) The Semantic similarity of the mapped sentences. Since we developed syntactic rules to cover the top 300 POS tags structures, the system coverage is to estimate how many sentences the system was
TABLE IX: A table showing the sentences that belong to the top 5 dominating POS tags structures (using chunking) after rewriting into a CNL.

<table>
<thead>
<tr>
<th>Rank</th>
<th>POS Pattern/Structure</th>
<th>Count</th>
<th>Examples</th>
<th>ACE CNL</th>
</tr>
</thead>
</table>
| 1    | COMP VBZ DT COMP IN   | 2664  | • Alkmene is a person in Greek mythology.  
      | COMP                  |       | • Anyang is a city in South Korea.  |
| 2    | COMP VBZ DT COMP      | 1399  | • Alicia Bridges is an American singer.  
      | VBZ                   |       | • Blood transfusion is a medical term.  |
| 3    | COMP VBZ DT COMP IN   | 1088  | • Gilbert is a city of Iowa in the United States.  
      | COMP IN DT COMP       |       | • Iphiclides is a genus of Butterflies in the family Papilionidae.  |
| 4    | COMP VBZ DT COMP IN   | 926   | • Anhui is a province in the east of China.  
      | DT COMP IN COMP       |       | • Agriculture is an important part of the economy of Azerbaijan.  |
| 5    | COMP VBD DT COMP      | 885   | • Abel Ricardo Laudonio was an Argentine boxer.  
      |                      |       | • Braniff International Airways was an American airline.  |

TABLE X: Evaluation of the coverage of the rewriting system.

<table>
<thead>
<tr>
<th>Metric</th>
<th>Total No. of sentences</th>
<th>Coverage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rewritten into ACE CNL</td>
<td>17,199</td>
<td>89.4%</td>
</tr>
<tr>
<td>Not rewritten into ACE CNL</td>
<td>2,037</td>
<td>10.6%</td>
</tr>
</tbody>
</table>

The resource is available for download and includes all the 17,199 sentences from original Simple Wikipedia and their parallel ACE rewritten sentences. These ACE sentences should be a very valuable resource to the community for exploitation in different applications related to NLP for knowledge base population.

V. CONCLUSION

We have presented a linguistic resource that is both human-readable and semantically machine interpretable. To our knowledge this resource is the first aligned dataset across a human-oriented and machine-oriented CNL as well as unstructured text. The dataset is well represented in that it takes almost the entire Simplified Wikipedia abstract population post-cleansing. We have provided corpus statistics and linguistics analysis, which have confirmed our hypotheses with respect to rewriting Simplified English to ACE and its subsequent transformation into logical and knowledge representations such as DRS and OWL respectively.

This resource could be exploited in other fields beyond CNLs for knowledge based population. Potential applications include generating general knowledge for expert and knowledge based systems and ontology aware NLP applications.

The resource is available for download and includes all the 17,199 sentences from original Simple Wikipedia and their parallel ACE rewritten sentences. These ACE sentences should be a very valuable resource to the community for exploitation in different applications related to NLP for knowledge base population.

We have presented a linguistic resource that is both human-readable and semantically machine interpretable. To our knowledge this resource is the first aligned dataset across a human-oriented and machine-oriented CNL as well as unstructured text. The dataset is well represented in that it takes almost the entire Simplified Wikipedia abstract population post-cleansing. We have provided corpus statistics and linguistics analysis, which have confirmed our hypotheses with respect to rewriting Simplified English to ACE and its subsequent transformation into logical and knowledge representations such as DRS and OWL respectively.

This resource could be exploited in other fields beyond CNLs for knowledge based population. Potential applications include generating general knowledge for expert and knowledge based systems and ontology aware NLP applications.

9https://drive.google.com/open?id=1eBUXZ8tESIML3jEptqG4aci4_ BnODkP
as well as knowledge based MT, automated reasoning, language learning as well as teaching and learning resources for knowledge engineering and logic programming. In addition to, generalization to other languages can be done using present frameworks such as GF to translate ACE to other languages [20].

Future work with respect to this resources will involve augmenting the aligned human to machine-oriented CNL content with semantic metadata such as RDF\textsuperscript{10} generated from ACE2OWL\textsuperscript{11} [17]. Other work will involve additional corpus analysis and rule generation for rewriting less common Simplified English POS patterns.

ACKNOWLEDGMENT

This work has emanated from research supported in part by a research grant from Science Foundation Ireland (SFI) under Grant Number SFI/12/RC/2289 and in part by the SSIX Horizon 2020 project (grant agreement No 645425).

REFERENCES


\textsuperscript{10}https://www.w3.org/RDF/
\textsuperscript{11}https://www.w3.org/2001/sw/wiki/ACE