Controlled Natural Language
B. Davis et al. (Eds.)

doi:10.3233/978-1-61499-904-1-85

Rewriting Simplified Text into a Controlled Natural Language

Hazem Safwat, Manel Zarrouk, Brian Davis
Insight Centre for Data Analytics,
National University of Ireland,
Galway, Ireland
{*hazem.abdelaal, manel.zarrouk, brian.davis*}*@insight-centre.org*

Abstract. While machine processable Controlled Natural Languages (CNLs) as a natural language interface have proven a popular, unambiguous and user friendly method for non experts to engineer formal knowledge-bases, human-oriented CNLs however remain under-researched despite having found favor within industry over many years. Whether such human orientated CNLs like the machine processable counterparts can be captured automatically as formal knowledge remains an open question. In addition, rewriting all or most of a human-oriented CNL into a machine-oriented CNL could unlock significant silos of general purpose domain knowledge, contained within existing human-oriented CNL content for exploitation by knowledge based systems. This paper explores the feasibility of rewriting a human-orientated CNL represented in Simplified English into a well know machine-oriented CNL represented in ACE CNL and describes preliminary results.

Keywords. Controlled Natural Language, Natural Language Processing, Knowledge Extraction, Semantic Web

1. Introduction

Knowledge extraction from text is still a problem that is not fully solved. There are a variety of advantages that makes solving this problem crucial in the community. Most of these advantages lie within the number of applications that can benefit from the extracted knowledge. These applications include but are not limited to generating quality linked data and ontologies from text, and generating machine-readable content to support Semantic Web Technologies. Although there are a lot of benefits, there are also many challenges with respect to the current approaches to ontology learning and population from text [1]. These limitations varies from the time needed to train non-expert users, to the time and effort needed to prepare the generated linked data to fit with different schemas and languages such as OWL. Some requirements presented in [2], are defined to help solve these problems. The requirements are related to mapping natural language to a well defined model representation (i.e. OWL), representing complex relations in text, and reduce or eliminate the need for training of non expert users for ontology engineering. However aside from explicit knowledge gathering and engineering activities machine orientated CNLs offer little incentive to the average user to create formal knowledge. A subcategory of CNL which offers a middle ground of reduced ambiguity

for language processing but less restriction than a machine-oriented CNL is a human-oriented CNL [8]. Their origins are motivated for the purposes of language learning, and unambiguous communication between humans in a domain context. An example of a human-oriented CNL is Simple English used to author Simple Wikipedia[1].

In this paper, we argue that rewriting a human orientated CNL (Simplified English) into a machine processable CNL allows us to reap the benefits of machine processable while simultaneously circumventing the barrier with respect to uptake of machine processable CNLs by users working outside of the knowledge engineering context. Finally, rewriting all or most of a human-oriented CNL content into a machine-oriented CNL could unlock significant rich silos of implicit general purpose formal knowledge, contained within existing human-oriented CNL content such as Simple Wikipedia.

The paper is structured as follows: Section 2 describes related work, Section 3 presents the approach, processing and analysis. In Section 4, we discuss some initial results with examples, and finally, Section 5 offers a conclusion.

2. Related Work

Attempto Controlled English (ACE) [3] is a widely known CNL that is mainly designed for knowledge representation. ACE texts are both human readable and machine processable that can be unambiguously mapped into different formal language such as Discourse Representation structures (DRS). The DRS is a variant of first order logic and its output of ACE texts can be bidirectionally translated into Web Ontology Language (OWL). Beside all the previous reasons, we choose ACE for testing our rewriting system as it provides open access to its tools and resources. One of these resources is the ACE parsing engine - APE[2], which we used for our system validation and for generating the DRS and OWL outputs.

The work from [5] proposed a sentence rewriting system based on semantic parsing that can rewrite a sentence into a new form while keeping the same target logical form. The rewriting system is trying to resolve the problem of vocabulary mismatch between natural language and ontology. This mismatch happens due to the various expressions in natural language and the fact that one can express the same meaning using different expressions and sentence structures. The system is supported by a ranking approach to select the best rewriting using the features of the semantic parsing and rewriting. Our work is similar to them as we also rewrite sentence into a new form that have the same target logical form. However, the difference is that our efforts are directed towards generating a machine processable text in the form of CNL that can be an alternative to the simplified text and can be mapped into triples.

In [4] the authors present an approach which is based on text rewriting for the aim of automatically generating labeled data that can be used for model training. The approach is implemented after analysing Simple Wikipedia and their parallel Wikipedia texts and extracting some rewrite rules. These rules are then used to produce different structures of sentences that are annotated with gold standard labels. Our work is similar to them as we used Simple Wikipedia and their parallel Wikipedia abstracts for analysing their texts and we used rewrite rules. However our analysis was performed to measure the overlap

[1] https://simple.wikipedia.org/wiki/

[2] https://github.com/Attempto/APE

between the properties of both texts and CNLs, and our rewrite rules are implemented to produce CNL alternatives of the simplified text for the aim of extracting OWL triples.

Finally, with respect to human-oriented CNLs, ASD Simplified Technical English[3] was developed to improve the readability and comprehensibility in technical documents. Another example is Boeing Technical English to improve the communication between people for air traffic control [9]. The development and planning of these CNLs appears often community driven, like Simplified Wikipedia, to have been based on Basic English [10].

3. Methodology

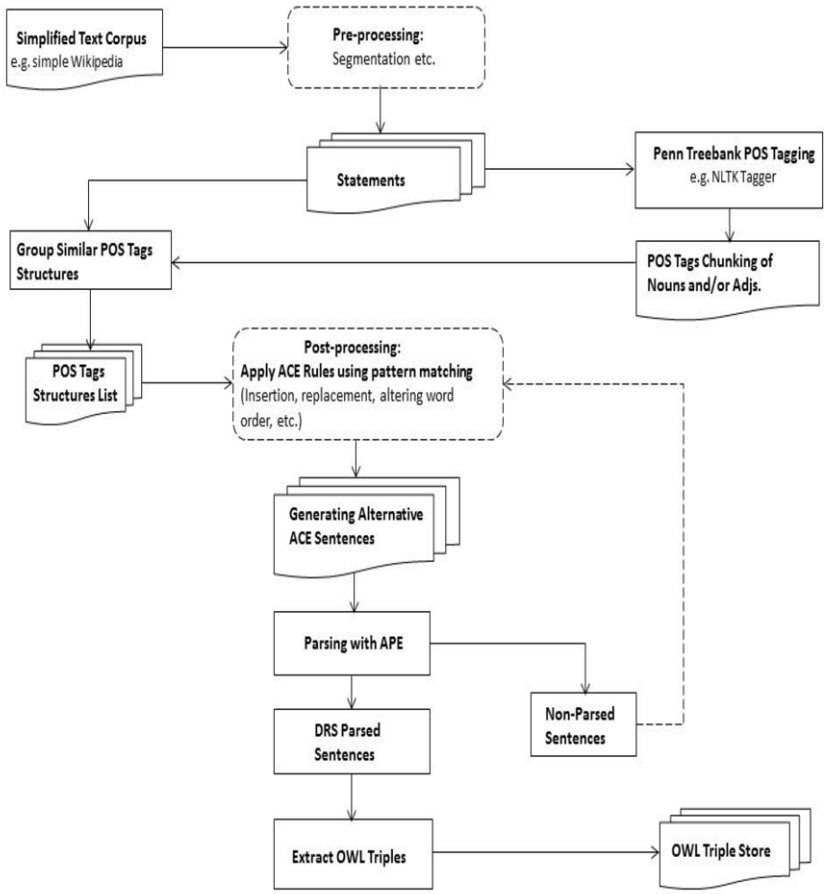

Figure 1. System architecture for rewriting simplified text into ACE CNL to extract OWL Triples.

The first step in our experiment involved collecting the dumps from both the Simple Wikipedia abstracts and its parallel Wikipedia abstracts. The main aim is to analyse the

[3] http://www.asd-ste100.org/

Table 1.: The most 5 common Chunked POS structures with an example of grouped POS structures (Descending) .

Chunked POS tags structures	Examples of grouped POS tags structures	Total No. of grouped POS tags structures	Total No. of sentences
CHUNK VBZ DT CHUNK IN CHUNK	• (NNP NNS) VBZ DT NN IN NN • (NNP NN) VBZ DT NN IN NNS	366	2664
CHUNK VBZ DT CHUNK IN DT CHUNK IN CHUNK	• NNP VBZ DT NN IN DT (JJ NN) IN NNP • (NNP NNP) VBZ DT NN IN DT NN IN NNP	224	926
DT CHUNK VBZ DT CHUNK IN CHUNK	• DT (NNP NNP) VBZ DT NN IN NNP • DT NNP VBZ DT NN IN (JJ NN)	194	469
CHUNK VBZ DT CHUNK IN DT CHUNK	• NNP VBZ DT NN IN DT (NNP NNP) • (NNP NNP) VBZ DT NN IN DT NNP	158	318
CHUNK VBZ DT CHUNK	• (NNP NNP) VBZ DT NN • NNP VBZ DT (JJ NN)	125	1399

text properties of each one and see if there is an overlap with the CNLs properties in [6]. The total number of sentences in Simple Wikipedia was 87k sentences and 39.2k in the parallel Wikipedia sentences. As shown from the system architecture in figure 1, text passes by a pre-processing module that involves text cleansing using regular expressions, then segmentation to split abstracts into sentences and tokenization to split sentences into tokens. The NLTK POS tagger that uses Penn Treebank [7] is applied on the sentences to analyse POS structures in each corpus. Our intuition is that, if the authors followed the guidelines in [6] to write the Simple Wikipedia abstracts, then there should be many similar POS tags structured sentences in the Simple Wikipedia corpus. From our analysis, we found that the sentences in the Simple Wikipedia abstracts overlap with the CNL properties more than the sentences written by authors in their parallel Wikipedia abstracts for instance the maximum tokens/sentence in the Simple Wikipedia corpus is always less than or equal 20 tokens as recommended in the CNL properties. In addition the use of passive voices is found only in approximately one third of the sentences, in contrary to the Wikipedia corpus where passive voices were found in more than half of the corpus.

The next step is extracting all the abstracts that follow the CNL rules defined in [6], which represent a total number of 36.5% sentences. Then, analyse the POS structures of the sentences to see to what extent the authors of Simple Wikipedia abstracts used the sentence forms such as `Subject-Verb-DirectObject`,

Table 2.: Examples of simplified text sentences after their rewriting into alternative ACE and generating their equivalent DRS and OWL outputs.

Simplified text sentences	Alternative ACE sentences	DRS Parser	OWL Parser
• Parametric statistics is a branch of statistics. • Herbs zoster is a disease in Humans.	• Parametric-statistics is a branch of Statistics. • Herbs-zoster is a disease in Humans.	Accepted	Accepted
• Lucca is a city in the Italian region of Tuscany. • Rio Cuarto is a city in the center of Argentina.	• Lucca is a city in the n:Italian-region of Tuscany. • Rio-Cuarto is a city in the center of Argentina.	Accepted	Accepted
• The Moscow Zoo is a zoo in Moscow. • The Manx is a breed of domestic cat.	• The n:Moscow-Zoo is a zoo in Moscow. • The Manx is a breed of the n:domestic-cat.	Accepted	Accepted
• Lubbock is a city in the United States. • North Holland is a province of the Netherlands.	• Lubbock is a city in the n:United-States. • North-Holland is a province of the Netherlands.	Accepted	Accepted
• Guilford County is a county. • Lyriel is a German band.	• Guilford-County is a county. • Lyriel is a n:German-band.	Accepted	Accepted

and `Subject-Verb-IndirectObject`[4]. The analysis showed that around 12.6k sentences can be grouped into 629 POS tags structures, which represent 34.5% from the total number of Simple Wikipedia sentences that can be mapped to ACE CNL. Hence we would need to implement rules that can cover the 629 different POS structures. We concluded that the percentage is not high, if we compared this to the total number of sentences which is 36.5k sentences. A deeper analysis is performed on the POS structures, where a large percentage from the POS structures could be grouped together as they contain noun and/or adjective clusters. So, after implementing a noun/adjective chunker, a new group of POS structures are created , where a total number of 19.2k sentences are grouped together under 300 POS structures.

From the previous analysis we can conclude that, if we can implement rules to rewrite the 300 POS tags structures into ACE CNL, then we can generate around 19.2k sentences into ACE CNL format and consequently into the DRS formal representation and exported to OWL Triples.

[4] `https://simple.wikipedia.org/wiki/Wikipedia:How_to_write_Simple_English_pages`

4. Initial Results

The system represented in the architecture shown in Figure 1 is implemented and applied on the 19.2k sentences that follow the CNL rules. In Table 1, we show the top five common chunked POS structures in the corpus, ranked in descending order by the total number of grouped POS structures under each chunked POS structure. The first column represents chunked POS structures, where the POS tag `CHUNK` refers to a one or more nouns and/or adjectives. Since, nouns can be found in different forms e.g singular, plural, noun phrases..etc, all of these forms are taken into consideration in the chunking process. The second column in the table shows a couple of examples of the original POS structures that are grouped under the chunked POS structure, with their total number shown in the third column. Finally, the last column shows the total number of sentences in the corpus that are grouped under this chunked POS structure.

Some examples that extend table 1 are shown in table 3, such that each row in table 3 is an extension of the same row in table 1. The *n:* prefix represents the chunked nouns performed by the chunker and the rewriting rules.

The first column in Table 3 contains some example Simplified Text sentences from the corpus. In the second column we show the sentences after implementing and applying the architecture in figure 1. The third and fourth columns, confirm the generated ACE alternatives are accepted and parsed by the ACE parser and the DRS and OWL outputs are generated, as shown in Table 3.

Table 3.: The DRS and OWL outputs from the ACE parser for an example sentence.

Metric	Result
Simplified text	Parametric statistics is a branch of statistics.
Equivalent ACE text	Parametric-statistics is a branch of Statistics.
DRS	[A,B]object(A,branch,countable,na,eq,1)-1/4 relation(A,of,named(Statistics))-1/5 predicate(B,be,named(Parametric-statistics),A)-1/2
OWL XML	$<Ontologyxml:base="http://www.w3.org/2002/07/owl\#"$ $xmlns="http://www.w3.org/2002/07/owl\#"$ $ontologyIRI="http://attempto.ifi.uzh.ch/ontologies/owlswrl/test">$ $<ObjectPropertyAssertion>$ $<ObjectPropertyIRI="http://attempto.ifi.uzh.ch/ontologies /owlswrl/test\#branch"/>$ $<NamedIndividualIRI="http://attempto.ifi.uzh.ch/ontologies /owlswrl/test\#Statistics"/>$ $<NamedIndividualIRI="http://attempto.ifi.uzh.ch/ontologies /owlswrl/test\#Parametric-statistics"/>$ $</ObjectPropertyAssertion></Ontology>$

5. Conclusion

This paper presented some initial results of our work towards rewriting simplified text (Simple Wikipedia) into a human-readable and machine-processable text (ACE CNL). Our initial results showed that, the features of simplified text are close to the features of CNL than unstructured text when users follow the authoring guidelines. We showed also that simplified text can be rewritten into a machine processable format (CNL) and can be used for knowledge extraction by extracting DRS and OWL triples. The approach

could be exploited to generate formal background knowledge for variety of applications automated reasoning, decision support or ontology aware NLP applications. The next steps will be applying the approach on all the abstracts of Simple Wikipedia to generate DRS and OWL triples, where this extracted structured knowledge could also be linked to a knowledge base such as DBpedia[5].

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

[1] Wong, Wilson, Wei Liu, and Mohammed Bennamoun. "Ontology learning from text: A look back and into the future." ACM Computing Surveys (CSUR) 44.4 (2012).

[2] Draicchio, Francesco, Aldo Gangemi, Valentina Presutti, and Andrea Giovanni Nuzzolese. "Fred: From natural language text to rdf and owl in one click." Extended Semantic Web Conference. Springer, Berlin, Heidelberg, 2013.

[3] Fuchs, Norbert E., Kaarel Kaljurand, and Tobias Kuhn. "Attempto controlled english for knowledge representation." Reasoning Web. Springer, Berlin, Heidelberg, 2008. 104-124.APA

[4] Woodsend, Kristian, and Mirella Lapata. "Text rewriting improves semantic role labeling." Journal of Artificial Intelligence Research 51 (2014): 133-164.

[5] Chen, Bo, et al. "Sentence rewriting for semantic parsing." Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Vol. 1. 2016.

[6] OBrien, Sharon. "Controlling controlled english. an analysis of several controlled language rule sets." Proceedings of EAMT-CLAW 3 (2003): 105-114.

[7] Marcus, Mitchell P., Mary Ann Marcinkiewicz, and Beatrice Santorini. "Building a large annotated corpus of English: The Penn Treebank." Computational linguistics 19.2 (1993): 313-330.

[8] Hujisen WO. 1998. *Controlled language: an introduction*. Second International Workshop on Controlled Language Applications (CLAW), Pittsburgh.

[9] Wojcik, Richard H., Heather Holmback, and James Hoard. 1998. *Boeing Technical English: An extension of AECMA SE beyond the aircraft maintenance domain*. Second International Workshop on Controlled Language Applications (CLAW), Pittsburgh.

[10] Ogden, Charles Kay. 1944. *Basic English: A general introduction with rules and grammar*. Vol. 29. K. Paul, Trench, Trubner.

[5] `http://wiki.dbpedia.org/`