A Rule-Based Approach for Converting Wikipedia Content into Semantic Relations

Nihal Yağmur Aydın[1], Tunga Güngör[2*]

[1] Boğaziçi University, Computer Engineering Dept., Istanbul, Turkey (e-mail: yagmur.aydin@boun.edu.tr).
[2] Boğaziçi University, Computer Engineering Dept., Istanbul, Turkey (e-mail: gungort@boun.edu.tr).

* Corresponding author. Tel.: +90 (212) 359 7094; email: gungort@boun.edu.tr
Manuscript submitted July 18, 2016; accepted August 30, 2016.
doi: ???

Abstract: In this paper, we propose a method for conversion from natural language into semantic relations. In this research, we focus on text written in highly unstructured form. The tools Stanford named entity recognizer and dependency parser are used to extract relevant information. The method is based on analysis of grammatical patterns of the sentences chosen from Wikipedia. The parse trees of sentences are examined in order to create patterns. Regular expressions are used to fetch the related nodes of the parse trees. In addition to the grammatical structure of sentences, we also made use of the named entities to create semantic relations. Experiments on different types of relations showed that 71% and 82% success rates can be obtained for a threshold of 0.50 correctness rate.

Key words: Grammatical patterns, Information extraction, Semantic representation, Wikipedia.

1. Introduction

Converting natural language texts into formal specifications is a quite challenging task that has importance in areas such as formal verification and model checking. This task also concerns with issues related to natural language understanding, information extraction, and question answering. There are studies about extracting formal specifications from structured text written in a proper format, such as getting formal verification properties from natural language documentation for HDL comments [1]. However, such studies focus on conversion from sentences written in highly structured forms by using syntactical properties of sentences.

In some other studies, the conversion from natural language into semantic relations was done by using outputs of a dependency parser [2]. SPARQL queries were run on the semantic relations. Some studies approach the conversion problem as a whole-sentence machine translation problem [3]. The translation process was considered as formed of five steps: rule extraction, local feature extraction, language model calculation, decoding, and tuning. Stanford named entity recognizer was used also for labeling information about location, person, etc.

In a research study on knowledge representation [4], XML was examined where ontologies were combined to add semantical representation of knowledge. In that approach, authors start with the XML schema. Afterwards, they map elements of XML schema into OWL language. Some definitions in XML data become objects, some of them become new relations, and some become attributes in that mapping. Lastly, reasoning tasks on XML schema are extended for ontologies.

In another study, two level grammars (TLG) were used to convert natural language to VDM++ specification [5]. Input was chosen as a data type, declaration, rule, rule statement, or meta sentence which contains information about the classifier, or a set of rules. Knowledge base was translated into TLG and then VDM++ specifications.

In embedded systems, ensuring correctness is of high importance. Therefore, operations regarding model

checking is crucial. For that purpose, property checking is applied to address the issue by extracting properties from the specification in terms of temporal logic expressions which can be subsequently checked by using model checker algorithms. In this respect, Wordnet, Stanford Dependency Parser and UML can be combined to make model checking [6].

Abstract Syntax Trees (ASTs) are generally used as intermediate representations for compiling high level language code into machine code. In a study for extracting formal specification from natural language [7], internal nodes of ASTs are chosen as operators (predicates), the subtrees that they dominate become the operands (arguments), and leaf nodes correspond to variables or constants.

ARSENAL [8] is a system used for text conversion and reasoning. In this system, relations extracted from Stanford dependency parser are used for TTEthernet requirements document. Intermediate representation table is formed of information like events, numericals, etc. ARSENAL first creates a graph. Each node is a mention entry and each (directed) edge indicates whether a mention is related to others via relations.

Conversion issue has also been used for deriving behavior specifications from textual use cases [9]. In another study focusing on formal verification of digital circuits using English specifications [10], sembolic model verification (SMV) model checker was used to get inferrential information from the text written in computation tree logic (CTL). The system consists of four components: (i) a parser, (ii) a convertor from semantic representations to CTL, (iii) the SMV model checker, and (iv) a module that mediates interaction between the three others.

Translation of natural language to OCL was performed in a study where the input text is natural language specification of an OCL constraint for a UML class model [11]. Sentence splitting, tokenization, POS tagging, lemmaziation (morphological analysis) are the first steps of the conversion. Stanford parser was used for conversion purposes. Mining text is closely related to extracting concepts from documents. For that purpose, in a study concept maps were created and extracted on the experiments done on short texts [12]. Generally, the subject of sentences represents the concept. Verbal phrase of the sentence is the object, representing a second concept. Relationship between subject and object were identified by the main verb in the sentence.

In this work, we focus mainly on conversion of sentences written in natural language into semantic relations using a rule-based approach. The novelty of the work originates from processing unstructured Wikipedia sentences in a particular domain. In addition to this, we combine the extracted grammatical patterns with named entity recognition outputs, which is not common in studies aiming at conversion of natural language sentences.

The paper is organized as follows: Section 2 describes the criteria for selection of sentences. In Section 3, we describe the grammatical patterns and semantic relations that are created. In Section 4, we discuss the results of the experiments. Finally, in Section 5, we conclude the work.

2. Sentence Selection

In this paper, we work on the domain of country information in Wikipedia. After a detailed analysis of country pages, we chose two types of sentences: i) Sentences revealing country specific information; ii) Sentences having general information regarding well-known people or brand names.

2.1. Country Specific Information

By examining the sentences in Wikipedia for country related data, we chose sentences having information regarding membership, area, population, climate, republic, state, border, location, economy, religion, and geographical coordinates. The relations and example sentences are listed in Table 1.

2.2. General Information

By analyzing named entity type of information on Wikipedia country pages, author, artist, composer, physicist, mathematician, and brand names were identified and they have been used for creating relations. The named entity relations and example sentences are listed in Table 2.

Table 1. Relations and Example Sentences for Country Specific Information

Relation	Example Sentence
border	Turkey is bordered by eight countries: Syria and Iraq to the south; Iran, Armenia, and the Azerbaijani exclave of Nakhchivan to the east; Georgia to the northeast; Bulgaria to the northwest; and Greece to the west.
area	With a territory of 110,994 square kilometers (42,855 sq mi), Bulgaria is Europe's 16th-largest country.
member	Germany is a member of UN, NATO, the G8, the G20 and the OECD.
population	The Netherlands had an estimated population of 16,785,403 on 30 April 2013.
location	Italy is a unitary parliamentary republic in Europe.
economy	The Netherlands has a market-based mixed economy, ranking 17th of 177 countries according to the Index of Economic Freedom.
republic	Bulgaria is a unitary parliamentary republic with a high degree of political, administrative, and economic centralization.
state	Russia is a sovereign state in northern Eurasia.
religion	Christianity is currently the largest religion in the Netherlands, accounting for about one-third of the population.
climate	Ukraine has a mostly temperate continental climate, although the southern coast has a humid subtropical climate.
geo	Egypt lies primarily between latitudes 22° and 32°N, and longitudes 25° and 35°E.

Table 2. Relations and Example Sentences for General Information

Relation	Example Sentence
author	Well-known German authors include Johann Wolfgang von Goethe, Friedrich Schiller, Gotthold Ephraim Lessing and Theodor Fontane.
composer	In the 19th century the most popular composers were: Józef Elsner and his pupils Fryderyk Chopin and Ignacy Dobrzyński.
artist	Distinguished contemporary artists include Roman Opałka, Leon Tarasewicz, Jerzy Nowosielski, Wojciech Siudmak, Mirosław Bałka, and Katarzyna Kozyra and Zbigniew Wąsiel in the younger generation.
physicist	Notable German physicists before the 20th century include Hermann von Helmholtz, Joseph von Fraunhofer and Gabriel Daniel Fahrenheit, among others.
mathematician	Numerous mathematicians were born in Germany, including Carl Friedrich Gauss, David Hilbert, Bernhard Riemann, Gottfried Leibniz, Karl Weierstrass, Hermann Weyl and Felix Klein.
brand	The new car market is dominated by domestic brands such as Renault (27% of cars sold in France in 2003), Peugeot (20.1%) and Citroën (13.5%).

3. Grammatical Patterns and Semantic Relations

In order to identify the grammatical patterns implicit in the text, we made use of the Stanford dependency parser [13]. Stanford dependency parser gives grammatical structure of sentences in the forms of trees. Grammatical categories such as noun, verb, or adjective are tagged with the related abbreviations on the parse tree. Due to that reason, there had been a need to extract words which had been associated with the tags. In order to achieve that, rules are generated by using the Tregex library of Java, which is a library to fetch the related nodes of a tree for the creation of relations [14]. Tregex library allows extraction of nodes based on grammatical patterns and regular expressions. Grammatical patterns which are used for the creation of relations are chosen by examining the parse tree generated by the Stanford dependency parser. Table 3 shows the grammatical patterns for relations and their meanings.

Then the relations are converted into logical representations using the Named Entity Recognizer (NER) [15] in addition to the parse of the sentences. NER tags the words in sentences as person, organization, or location. Therefore, by using NER it had been possible to extract meaningful information for the creation of relations. Table 4 shows the usage of named entity recognition in relations as well as their abstract representations formed by combining it with grammatical patterns. As shown in the table, combination of grammatical patterns with outputs of the named entity recognizer has resulted in generation of relations with unary and binary arity.

Table 3. Grammatical Patterns

Relation	Grammatical Pattern	Meaning
area	NP < CD	NP immediately dominates CD
population	NP < CD	NP immediately dominates CD
location	PP < NP	PP immediately dominates NP
economy	NP < JJ & << NN	NP immediately dominates JJ and NN dominates them
republic	NP < JJ & << NN	NP immediately dominates JJ and NN dominates them
state	NP < JJ & << NN	NP immediately dominates JJ and NN dominates them
religion	NP < NP	NP immediately dominates NP
climate	NP < JJ & << NN	NP immediately dominates JJ and NN dominates them
geo	CD	CD is taken from the sentence

Table 4. Abstract Representation of Relations

Relation	NER	Abstract Representation	Example Relation
border	Location	border (Location, pattern)	border (Turkey, Bulgaria)
area	Location	area (Location, pattern)	area (Italy, 301,338)
member	Organization	member (Organization, pattern)	member (Germany, UN)
population	Location	population (Location, pattern)	population (Bulgaria, 7,364,570)
location	Location	location (Location, Location)	location (Greece, Europe)
economy	Location	economy (Location, pattern)	economy (Netherlands, market-based)
republic	Location	republic (Location, pattern)	republic (Bulgaria, parliamentary)
state	Location	state (Location, pattern)	state (France, sovereign)
religion	Location	religion (Location, pattern)	religion (Netherlands, Christianity)
climate	Location	climate (Location, pattern)	climate (Denmark, temperate)
geo	Location	geo (Location, pattern)	geo (Egypt, 32N)
author	Person	author (Person)	author (Goethe)
composer	Person	composer (Person)	composer (Bach)
artist	Person	artist (Person)	artist (Boucher)
physicist	Person	physicist (Person)	physicist (Helmholtz)
mathematician	Person	mathematician (Person)	mathematician (Galilei)
brand	Organization	brand (Organization)	brand (Peugeot)

4. Experiments and Results

In order to evaluate the proposed method, we tested each relation with about 10 sentences compiled from country pages. Sentences containing keywords related to the relation name were chosen in the first step. Then, the success of the system was measured based on the number of sentences and their outputs. Table 5 shows the results and the analysis of different types of errors. The explanations of the columns in the table are as follows:

- A: The result is correct as depicted in Table 4.
- B: The result is partly correct. One of the arguments in the representation is not the correct answer, but the correct answer can be inferred.
- C: The result is incorrect. The error in the result is due to incorrect named entity output by the Stanford dependency parser.
- D: The result is incorrect. The error in the result is due to incorrect parse output by the Stanford dependency parser.
- E: The result is incorrect. The grammatical pattern (Table 3) is not applicable for the sentence.
- F: Total number of sentences for the relation (i.e. A+B+C+D+E).
- Correct-1: A/F
- Correct-2: A/(F-C-D)

The first correctness measure (Correct-1) shows the accuracy when all types of errors (B,C,D,E) are taken into account. However, the errors denoted by the columns C and D originate from the incorrect outputs of the Stanford tools, on which the approach in this paper is based. Therefore, we give an additional correctness measure (Correct-2), in which the sentences parsed incorrectly are excluded.

Table 5. Evaluation Results and Analysis of Errors

Relation	A	B	C	D	E	F	Correct-1	Correct-2
border	3	3	4	0	0	10	0.30	0.50
area	6	1	0	1	2	10	0.60	0.66
member	9	0	0	1	0	10	0.90	1.00
population	4	0	0	2	4	10	0.40	0.50
location	7	0	0	1	2	10	0.70	0.78
economy	3	5	1	1	0	10	0.30	0.38
republic	10	0	0	0	0	10	1.00	1.00
state	8	0	0	2	0	10	0.80	1.00
religion	2	0	0	2	6	10	0.20	0.25
climate	4	0	1	0	5	10	0.40	0.44
geo	10	0	0	0	0	10	1.00	1.00
author	10	0	0	0	0	10	1.00	1.00
composer	10	0	0	0	0	10	1.00	1.00
artist	10	0	0	0	0	10	1.00	1.00
physicist	6	1	0	0	0	7	0.86	0.86
mathematician	3	2	0	0	0	5	0.60	0.60
brand	7	0	0	0	0	7	1.00	1.00

If we take 0.50 as a threshold for the success rate, we see that 12 out of 17 relations with respect to Correctness-1 (71%) and 14 out of 17 relations with respect to Correctness-2 (82%) are successful. Below we give a brief analysis for each relation type.

For the area relation, the problem related to the parser output is caused by the representation of area information as adjective (JJ) rather than cardinal (CD). An example sentence is the following: "*Italy covers an area of 301,338 km2 (116,347 sq mi) and has a largely temperate seasonal climate; due to its shape, it is often referred to in Italy as lo Stivale (the Boot).*". For the population relation, an example with grammatical pattern problem is: "*Egypt is the most populated country in the Middle East, and the third most populous on the African continent, with about 88 million inhabitants as of 2015.*" The output of this relation is given as "*population (Egypt, 2015)*".

For the republic and geo relations, the identified patterns were shown to be suitable for the sentence structures and thus all results were correct. For the economy relation, half of the results were partly correct, which is related to both patterns and sentence structures. An example sentence of economy relation is as follows: "*Italy has a capitalist mixed economy, ranking as the third-largest in the Eurozone and the eighth-largest in the world.*". This sentence resulted in the output "*economy (Italy, third-largest)*". For the religion and climate relations, it was quite challenging to generate a general rule since the sentences take highly different forms. Thus, a significant number of pattern errors was observed. For the member and state relations, errors were due to problems in the Stanford parser, which can be ignored for the evaluation. For the location relation, a small number of errors was caused by the pattern rule.

For the border relation, named entity problems occurred due to the mixed structure of sentences. An example of this is the relation "*border (Ukraine, bordered by Russia to)*", extracted from the sentence "*Ukraine bordered by Russia to the east and northeast, Belarus to the northwest, Poland and Slovakia to the west, Hungary, Romania, and Moldovato the southwest, and the Black Sea and Sea of Azov to the south and southeast, respectively.*". For the relations of author, composer, artist, brand, mathematician and physicist, no errors were detected.

5. Conclusions

In this work, a method for converting a natural language text into semantic relations was proposed. By using a dependency parser and named entity recognizer, 17 different relations were identified using sentences from Wikipedia documents. The success rates of the relations ranged between 20%-100%. It was shown for 71% of the relations more than half of the outputs were correct. When the errors related to the tools are ignored, the success rates ranged between 25%-100%, and percentage of relations above 0.50 threshold raised to 82%. As future work, relations on other domains or approaches based on syntactic trees can be considered.

References

[1] Harris, C. B., & Harris, I. G. (2015). Generating formal hardware verification properties from natural language documentation. In *IEEE conference on semantic computing* (pp. 9-56).

[2] Drechsler, R., Harris, C. B., Harris, I. G., Abdessaied, N., & Soeken, M. (2014). Automating the translation of assertions using natural language processing techniques. In *IEEE forum on specification and design languages* (pp. 1-8).

[3] Hermjakob, U., Knight., K., Marcu, D., May, J., & Pust, M. (2015). Parsing English into abstract meaning representation using syntax-based machine translation. In *Proceedings of EMNLP* (pp. 1143-1154).

[4] Cheng, J., Ma, Z. M., Yhan, L., & Zhang, F. (2011). Knowledge representation and reasoning of XML with ontology. In *Proceedings of ACM symposium on applied computing* (pp. 1705-1710).

[5] Lee, B., & Bryant, B. R. (2002). Contextual knowledge representation for requirements documents in natural language. In *Proceedings of FLAIRS*.

[6] Drechsler, R., Soeken, M., & Wille, R. (2012). Formal specification level: towards verification-driven design based on NLP. In *IEEE forum on specification and design languages (FDL)* (pp. 53-58).

[7] Dinesh, N., Joshi, A., Lee, I., & Webber, B. (2006). Extracting formal specifications from natural language regulatory documents. In *Proceedings of the international workshop on inference in computational semantics*.

[8] Elenius, D., Ghosh, S., Li, W., Lincoln, P., Shankar, N., & Steiner, W. (2014). ARSENAL: Automatically extracting requirements specifications from natural language. In *Proceedings of NASA formal methods symposium*.

[9] Mencl, V. (2004). Deriving behaviour specifications from textual use cases. In *Proceedings of international conference on software engineering* (pp. 235-244). ACM.

[10] Holt, A. (1999). Formal verification with natural language specifications: spectrum guidelines, experiments, lessons so far. *South African Computer Journal*, *24*, 253-257.

[11] Bajwa, I. S., Lee, M., & Bordbar, B. (2012). Translating natural language constraints to OCL. *Journal of Computer and Information Sciences*, *24(2),* 117-128.

[12] Wang, W. M., Cheung, C. F., Lee, W. B., & Kwok, S. K. (2008). Mining knowledge from natural language texts using fuzzy associated concept mapping. *ACM Journal of Information Processing and Management*, *44(5),* 1707-1719.

[13] Stanford dependency parser. Available: http://nlp.stanford.edu/software/lex-parser.shtml.

[14] Tregex. Available: http://nlp.stanford.edu/nlp/javadoc/javanlp/edu/stanford/nlp/trees/tregex/TregexPattern.html.

[15] Stanford named entity recognizer. Available: http://nlp.stanford.edu/ner/.

Nihal Yağmur Aydın received M.S. degree from Department of Computer Engineering, Boğaziçi University, Istanbul, Turkey, in 2016. She worked as a scholarship holder in national supported projects. Her research interests include natural language processing, pattern recognition, and text summarization.

Tunga Güngör received Ph.D. degree from Department of Computer Engineering, Boğaziçi University, Istanbul, Turkey, in 1995. He is currently a full professor in the Department of Computer Engineering, Boğaziçi University. His research interests include natural language processing, machine translation, machine learning, and pattern recognition. He published about 80 scientific articles, and participated in several research projects and conference organizations.