Parsing Indian English News Headlines

Samapika Roy[1], **Sukhada**[1], **Anil Kr. Singh**[2]
[1]Dept. of Humanistic Studies, IIT (BHU)
[2]Dept. of Computer Science and Engg., IIT (BHU)
{samapikar.rs.hss15, sukhada.hss, aksingh.cse}@itbhu.ac.in

Abstract

Parsing news headlines is one of the difficult tasks of Natural Language Processing (NLP). It is mostly because news headlines (NHs) are not complete grammatical sentences. News editors use all sorts of tricks to grab readers' attention. For instance, unusual capitalization as in headline 'Ear SHOT ashok rajagopalan'; some demand world knowledge like 'Church reformation celebrated' where 'Church reformation' refers to a historical event and not a piece of news about an ordinary church. The lack of transparency in NHs can be linguistic, cultural, social, or contextual. The lack of space provided for a news headline has led to creative liberty.

Though many works like news value extraction, summary generation, emotion classification of NHs have been going on, parsing them had been a tough challenge. Linguists have also been interested in NHs for creativity in the language used by bending traditional grammar rules. Researchers have conducted studies on news reportage, discourse analysis of NHs, and many more. While the creativity seen in NHs is fascinating for language researchers, it poses a computational challenge for NLP researchers. This paper presents an outline of the ongoing doctoral research on the parsing of Indian English NHs. The ultimate aim of this research is to provide a module that will generate correctly parsed NHs. The intention is to enhance the broad applicability of newspaper corpus for future NLP applications.

1 Introduction

NHs stands to be an excellent example of creative writing. Headlines tend to be short, attention-grabbing, and giving out just enough information to attract readers' attention. To keep readers engage, news editors use all sorts of tricks possible. Studies on Indian English have been going on for some decades and being a part of South Asian English, Indian English has managed to grab attention for the past few years. Like other stratification of the English language, Indian English is unique as well. Though it follows the basic underlying structure of British English, the syntax, semantics, and pronunciation of Indian English differ from world English. Translating from native language to English sometimes results in different sentence structures. Due to linguistic diversity, code-mixing and code-switching are extremely common. These effects can be found in NHs as well.

Sometimes news editors tend to use local idiosyncrasies. NHs are different from standard text. For example, we can find unusual capitalization as in 'Ear SHOT ashok rajagopalan', deliberate subject noun phrase drop as in 'Arrested for theft', the deliberate dropping of main verb as in 'Identity cards for all urban street vendors', dropping of auxiliaries as in '18th century stone inscription unearthed', so on. Such structures are unique to NHs which make it difficult to parse them using existing parsers. This research attempts to solve such problem of NHs parsing.

2 Contributions

This research contributes at following levels :

1. NHs Corpus: A corpus of 40,000 (approx.) headlines containing 3 lakh words (approx.) of Indian English NHs has been collected.
2. Parallel corpus: A parallel corpus of NHs and grammatically transformed corresponding sentences has been created.
3. Linguistic analysis of the NHs data: Through the Linguistic analysis, different structures of NHs, words compositions, voices, tenses, dropping of subjects, as well as usage of punctuation have been observed.

Proceedings of the 17th International Conference on Natural Language Processing, pages 239–242
Patna, India, December 18 - 21, 2020.

4. Guideline creation: A proper set of guidelines for the transformation of NHs has been drafted covering the various structures of NHs that were found out as a result of linguistic analysis,
5. Feature model: We have created a syntactico-semantic feature model based on linguistic analysis conducted on NHs corpus.
6. Headline grammar: The creation of news headline grammar based on the linguistic analysis is going on.
7. NHs Module: Creation of a module which could provide us correct parsing of NHs is in the process.

3 Methodology

For the study, we deliberately chose Indian English and collected data for two reasons: 1) apart from the linguistic analysis, a contrastive analysis between NHs of Indian English and British English has been conducted. This study helped us to understand whether the structure of NHs in Indian English is different from NHs in British English and other English varieties, 2) Working on Indian English, collecting Indian English NHs data and building a parallel corpus will be fruitful for future endeavors as some computational works have been already conducted on Indian English. On the other hand, Indian English data collection is going on for quite a some time in many organizations.

We collected data from three different newspapers to analyze the syntactic structures. The lack of space in hard copy newspapers leads to an opportunity for creative liberty for news editors, like excluding grammatical and lexical elements (dropping subject noun phrase, auxiliary drop), the-out-of-grammar use of punctuation, and so on. For a systematic analysis of the data, we adhered to an inductive research strategy.

Our very first objective was to analyze the data for a better understanding of the problem. This analysis helped us to come up with an idea about accuracy of the currently available parsers and areas where parsing output can be improved. After parsing (Constituency) the data with existing open-source parsers like Stanford and Allennlp, we observed several errors like parsing of singular verbs as plural nouns, adjectives as verbs, and so on. There was a need to understand the structures of NHs and thus to analyze the NHs corpus linguistically. This study was exploratory and interpretive. For this we used both qualitative and quantitative methods for the data analysis.

We found out that NHs can be either complete headlines (headlines following the necessary SVO word order of English language) and fragments (comprises mostly of noun phrases, ex. 'A burning issue'). The complete NHs were further divided into four categories according to structures: declarative, imperative, interrogative, and exclamatory. The grammatical notions we used to describe the headlines are in a broader sense as headlines can not fulfill every linguistic need to belong in any particular category. Instead, they fulfilled mere basic needs. The linguistic analysis of NHs helped us create a grammar and guideline to transform NHs to their canonical form to get the correct parse.

Based on the linguistic analysis, we have build a syntax and semantics-based feature model. In this model, we have tried to cover the various headlines' structures we have come across so far in our study. It has been created with the motive that it will help us in the annotation of the NHs. Also, it will work as a guideline to create a transfer grammar analysis module. In the feature model, we covered all the aspects of NHs like headlines structure, which can be a complete NH or a fragment; there are NHs with or without subject noun phrase; NHs which is only a quote from a speaker but without the mention of the speaker, and so on. Our next step is to automate the transformation. We decided to treat the problem as a machine translation problem, for which we are working on creating a parallel corpus of raw NHs and equivalent grammatically transformed sentences. We are also working on creating transfer grammar for this purpose.

4 Observations

We have observed following specific vital issues during the research so far:

1. Incorrect parsing: We observed that parsing the NHs with the current available parsers is difficult as some grammatical elements are intentionally dropped in NHs. The error analysis helped us to identify the structural issues of parsing the NHs with existing parsers. In order to understand the syntax of NHs, the linguistic analysis of the NHs corpus became necessary. For example, Nouns marked as Verbs: (ROOT (S (NP (CD Two) (NN policemen)) (VP (VBD suspended) (SBAR (IN as) (S (VP (VBD

accused) **(NP (NNS escapes)**) (PP (IN from) (NP (NN custody)))))))))

2. Abridged structure: The linguistic analysis of the NHs gave us an insight into the internal structure of NHs which leads to a grammar which is specific to headlines.
3. Limited information: We faced specific issues while transforming the NHs into grammatical sentences. For transformations, our intention was not to add anything that is not certain from the NHs as well as be careful in not removing any vital information from the NHs. With the limited information we can get from the NHs, we focused on fulfilling the basic yet essential grammatical requirements.
4. Approach: Since there is no existing framework from which we can draw inspirations, we decided to go with hybrid approach in creating our module for parsing the NHs.

5 Results

The results of the error analysis of the data after constituency parsing, shows that incorrect tags e.g. plural nouns as an adjective, singular verbs as plural nouns, common nouns as proper nouns and so on led to incorrect parsed output. Following is an example of such output which shows multiple parsing errors: (ROOT(NP(NP **(NNP Boat)**)(NP **(JJ capsize)** (NN toll) **(NNS touches)**) (NP (CD 21)))). As we can observe here, 'Boat' has been incorrectly parsed as NNP(proper noun)'capsize' above is incorrectly parsed as JJ (adjective) and 'touches', above is incorrectly tagged as NNS (singular noun). The reason behind these incorrect tags is the intentional dropping of grammatical elements by editors due to space constraints. The errors occurred mostly because these NHs are not grammatically structured. The errors are the lack of crucial grammatical elements in the NHs, which distorts the grammatical bond in a sentence.

To cross-examine our observations, we provided the parsers with sample data of grammatically transformed raw NHs, where we added a few essential grammatical elements required. Proving our theory (which we developed from error analysis) correct, the parsers provided the correct parse. This finding has important implications for developing the module we intend for the generation of correctly parsed NHs. In linguistic analysis, we observed specific structural constructions like dropping off the subject and out of the grammar style of using linguistic devices like punctuation, idioms, multiword expressions, and many more.

6 Future Goals

Future research will be devoted to the development of news headline grammar and on the viability of our approach. We have already conducted linguistic analysis, which will act as a framework for headline grammar. We are currently working on automating the NHs corpus to the standard canonical sentences. We intend to elaborate the research on the contrastive analysis between a parallel corpus of raw headlines and its grammatically transformed sentences. For the rule-based approach, we are moving forward with the creation of context-free-grammar rules. For automating the NHs to its grammatical form, we are considering LALR parsers to start. In the meantime, we are trying to create enough parallel corpus of the raw NHs corpus we collected and the grammatically transformed sentences of those NHs to go with a statistical approach.

7 Research Roadmap

The research road map is as follows: We did data collection and sanitization during this research followed by constituency parsing of data, performed the error analysis, conducted the linguistic analysis of NHs, did manual transformations of NHs for the rule-based approach, and formulated a guideline for the transformations, and build a feature design model for the NHs. We are currently working on constructing a news headline grammar, automating the feature annotation process, and finally building a module to provide us with correctly parsed NHs as final output.

Acknowledgments

I would like to express my sincere gratitude to my supervisor Dr. Sukhada, Assistant Professor, dept. of Humanistic Studies, IIT (BHU) and my co-supervisor Dr. A.K. Singh, Associate professor, dept. of Computer Science and Engg., IIT (BHU) for their constant support and guidance.

References

Innocent Ejimofor Agu. 2015. A linguistic-stylistic analysis of newspaper reportage. *International Journal*, 20.

Alireza Bonyadi and Moses Samuel. 2013. Headlines in newspaper editorials: A contrastive study. *Sage Open*, 3(2):2158244013494863.

Christine Develotte and Elizabeth Rechniewski. 2001. Discourse analysis of newspaper headlines: a methodological framework for research into national representations. *The Web Journal of French Media Studies*, 4(1):1–12.

Daria Lombardi. 2018. Critical discourse analysis of online news headlines: A case of the stoneman douglas high school shooting.

Novriyanto Napu. 2018. English and indonesian newspaper headlines: A comparative study of lexical features. *European Journal of Literature, Language and Linguistics Studies*.

Alicja Piotrkowicz, VG Dimitrova, and Katja Markert. 2017. Automatic extraction of news values from headline text. In *Proceedings of the Student Research Workshop at the 15th Conference of the European Chapter of the Association for Computational Linguistics (EACL SRW 2017)*, pages 64–74. Association for Computational Linguistics.

Satoru Takahashi, Masakazu Takahashi, Hiroshi Takahashi, and Kazuhiko Tsuda. 2007. Analysis of the relation between stock price returns and headline news using text categorization. In *International Conference on Knowledge-Based and Intelligent Information and Engineering Systems*, pages 1339–1345. Springer.