Morphological Analysis of Compound Words in Hindi Language

Abhigya Agrawal

October 18, 2022

1 Abstract

In linguistics, morphological analysis is a process used to break down words into morphemes. Morphemes can be words or affixes that add meaning to the word. Compound words are words formed by combining two or more words. What makes compound words interesting is that they can have a completely different meaning from their source words. For example butterfly is a creature with wings and not a fly made of butter. The morphological analysis of compound words is the process of breaking these words into the stem words and categorizing them. Some researchers like Chakrabarti et al. (2008), Deepa et al. (2004), Dasgupta et al. (2010), Dwivedi and Ghosh (2022), etc. have been intrigued with compound words in Hindi and other Indian languages and have provided the ground work for this project. Work done by Lang et al. (2022) and Dwivedi and Ghosh (2021) has provided some insight into how to proceed with the morphological analysis.

This project will make use of both linguistic methods such as lexicon as well as machine learning models to successfully carry out morphological analysis of Hindi compound words. The initial task of the project is to prepare the data, the lexicon and decide the categories the words will be classified into. The purpose of morphological analysis is to get a better understanding of compound words and therefore get a better understanding of language modeling. This also aids other language based tasks such as

speech recognition, word to text models and AI assistants like Alexa.

2 Dataset

Singh et al. (2016) have created a multi-word expressions dataset for Indian languages. A subset of this dataset focuses on compound nouns in Hindi language. The dataset consists of about 12000 part-of-speech-tagged compound nouns and can be downloaded from `https://www.cfilt.iitb.ac.in/download_new.html`.

3 Evaluation

Since the core task of this project is classification, the evaluation will be done using accuracy, F1-macro score, F1-micro score, precision and recall.

References

Debasri Chakrabarti, Hemang Mandalia, Ritwik Priya, Vaijayanthi Sarma, and Pushpak Bhattacharyya. Hindi compound verbs and their automatic extraction. In *Coling 2008: Companion volume: Posters*, pages 27–30, Manchester, UK, August 2008. Coling 2008 Organizing Committee. URL https://aclanthology.org/C08-2007.

Sajib Dasgupta, Naira Khan, Asif Sarkar, Shahriar Hossain, and Mumit Khan. Morphological analysis of inflecting compound words in bangla. 09 2010.

S.R. Deepa, Kalika Bali, A.G. Ramakrishnan, and Partha Pratim Talukdar. Automatic generation of compound word lexicon for Hindi speech synthesis. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*, Lisbon, Portugal, May 2004. European Language Resources Association (ELRA). URL http://www.lrec-conf.org/proceedings/lrec2004/pdf/501.pdf.

Vandana Dwivedi and Sanjukta Ghosh. Classification of hindi compound nouns using machine learning - sn computer science, Oct 2021. URL https://link.springer.com/article/10.1007/s42979-021-00895-z.

Vandana Dwivedi and Sanjukta Ghosh. Interpretation of hindi compound nouns using word2vec embeddings, May 2022. URL https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4121850.

Inga Lang, Lonneke Plas, Malvina Nissim, and Albert Gatt. Visually grounded interpretation of noun-noun compounds in English. In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, pages 23–35, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.cmcl-1.3. URL https://aclanthology.org/2022.cmcl-1.3.

Dhirendra Singh, Sudha Bhingardive, and Pushpak Bhattacharyya. Multiword expressions dataset for Indian languages. In *Proceedings of the*

Tenth International Conference on Language Resources and Evaluation (LREC'16), pages 2331–2335, Portorož, Slovenia, May 2016. European Language Resources Association (ELRA). URL https://aclanthology.org/L16-1369.