Identifying Support Words for Nominal Predicates

Aashka Trivedi (aht323), Raksha Hegde (rh3468), Sarvani Nadiminty (sn2884)

7 April 2022

This project aims to identify the *support* words (words that connect the argument to the predicate) for nominal predicates from Nombank [1]. Here, we only use sentences from Nombank that contain a word with a *support* label, in addition to the *predicate* and *arguments*, and include all types of nominal predicates. Initial preprocessing results in over 2000 training sentences, 1000 test sentences, and 600 development sentences containing *support* words. We will experiment with the use of support verbs and transparent nouns as additional features, or means to obtain more sentences for training or testing purposes.

Since the work involving the identification of support words for Nombank is relatively sparse, we hope to provide insights into feature selection and relative performance of ML Algorithms for this task. However, we draw inspiration from prior work involving the automatic identification of *light verbs* or *support verbs* [2, 3, 4, 5], which lack semantic meaning in themselves, but are usually ejected from the original verb's syntactic role during the nominalization. Halliday's[6] notion of *Lexical Density* where each word carries different semantic weight has been the base to identify the semantic lightness of any word. For example, the verb *propose* is nominalized int *make a proposal*, where *make* is the corresponding support or light verb making it semantically light. Our model will identify light verbs with Noun Phrase complements only, by studying the link between a verb's relative frequency in these light verb expressions and its semantic lightness.

Grefenstette and Teufel [2] agree that the support verb to be used with a nominalized predicate structure is unpredictable. However, they introduce a method of deriving probable support verbs given a verb and its nominalization, using POS tagging, syntactic analysis using a robust parser and frequency statistics over a corpora. Taking all verb-object pairs (object is deverbal nominalisation) is one technique to extract light verbs from a corpus. Grefenstette and Teufel [2] use only Local Information (specific nominal)- when all the occurrences of noun in verb-object pairs are counted, a local relative frequency for each verb is calculated with regard to that noun. Dras [3] uses Local information with Global information. A global relative frequency for each verb is calculated by counting all instances of that verb, independent of their objects. Both local and global information are integrated to generate a changed likelihood of becoming a support verb. His model revealed some information on how sometimes there can be more than one proposed support verb. For ex: *cause/do harm*. Here both *cause* and *do* can be a good choice for a support verb and their frequency ratio is quite small (Frequency ratio = ratio of product of local and global relative frequencies for the first choice to the product of local and global relative frequencies second choices). This is true for many other support verbs like *bear/have resemblance* or *make/produce change*. It is also apparent that the relative frequency of a verb in light verb constructions appears to be a very reasonable predictor of a verb's supporting behaviour, resulting in viable options for support verbs for nominalizations.

Vincze et.al in [4] uses rule and Machine Learning based approaches to identify Light Verb Constructions(LVC) from Multiword expressions in different domains. POS-rule method accepts the n-grams for which the predefined patterns such as VB.? (NN—NNS) could be applied as LVC. The 'Most Frequent Verb'(MFV) method uses the fact that most of the light verbs are the most frequently occurring verbs (e.g. do, make, take etc.). Various other methods such as 'Suffix','Stem' etc. are combined with the ones above in the rule based approach. The MFV method and the intersection of MFV and 'Stem' was

found to be the most useful rule based approaches. In the Machine Learning approach, a Conditional Random Field(CRF) classifier is used along with features - domain specific dictionaries of first names, company types, token frequency, POS tags and contextual information such as sentence position, trigger words, etc. CRF classifier yields better results when used with an extended feature set that includes the rule-based methods as LVC specific features. The features used for LVC identification in [5] are POS tags of the support verb(Vs) and the Noun(N) pairs and the surrounding words, the dependency tree node distance between Vs and N, the voice (active/passive) of the Vs and the word sense tags from OntoNotes and WordNet. These can be further explored for our task.

The approach of this project will be incremental- we will first establish an upper-bound to the performance of our system by assuming that it is knowledgeable of both the predicate and the argument(s) of each sentence, and use features incorporating this knowledge to predict the support. We then move to the more difficult set-up, where the system only knows the predicate, and must jointly identify the argument and the support, perhaps in a way similar to the work of Zhao et. al [7], which uses a single system to handle all the sub-tasks of traditional Semantic Role Labelling, predicate identification/disambiguation and argument identification/classification. A lower bound would be a system having absolutely no information about the predicate, argument(s) or support- and must jointly predict all three roles. The intuition of this project would be to produce a system that can work reasonably well with different information levels.

References

[1] A. Meyers, R. Reeves, C. Macleod, R. Szekely, V. Zielinska, B. Young, and R. Grishman, "The NomBank project: An interim report," in *Proceedings of the Workshop Frontiers in Corpus Annotation at HLT-NAACL 2004*, (Boston, Massachusetts, USA), pp. 24–31, Association for Computational Linguistics, May 2 - May 7 2004.

[2] S. Teufel and G. Grefenstette, "Corpus-based method for automatic identification of support verbs for nominalizations," in *Seventh Conference of the European Chapter of the Association for Computational Linguistics*, (Dublin, Ireland), Association for Computational Linguistics, Mar. 1995.

[3] M. Dras, "Automatic identification of support verbs: A step towards a definition of semantic weight," in *Proceedings of the Eighth Australian Joint Conference on Artificial Intelligence*, (Singapore), pp. 451–458, World Scientific Publishing Co. Pte, 1995.

[4] I. T. Nagy, G. Berend, G. Móra, and V. Vincze, "Domain-dependent detection of light verb constructions," in *Proceedings of the Second Student Research Workshop associated with RANLP 2011*, (Hissar, Bulgaria), pp. 1–8, Association for Computational Linguistics, Sept. 2011.

[5] W.-T. Chen, C. Bonial, and M. Palmer, "English light verb construction identification using lexical knowledge," in *AAAI*, pp. 2368–2374, 2015.

[6] M. A. K. Halliday, "Spoken and written language," Waurn Ponds, Vic. : Deakin University : distributed by Deakin University Press, 1985.

[7] H. Zhao, W. Chen, and C. Kit, "Semantic dependency parsing of NomBank and PropBank: An efficient integrated approach via a large-scale feature selection," in *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, (Singapore), pp. 30–39, Association for Computational Linguistics, Aug. 2009.