Automatic Detection of Problematic Rules in Vietnamese Treebank

Hong-Quan Nguyen[1], Phuong-Thai Nguyen[2],
Thanh-Quyen Dang[3], Van-Hiep Nguyen[4]

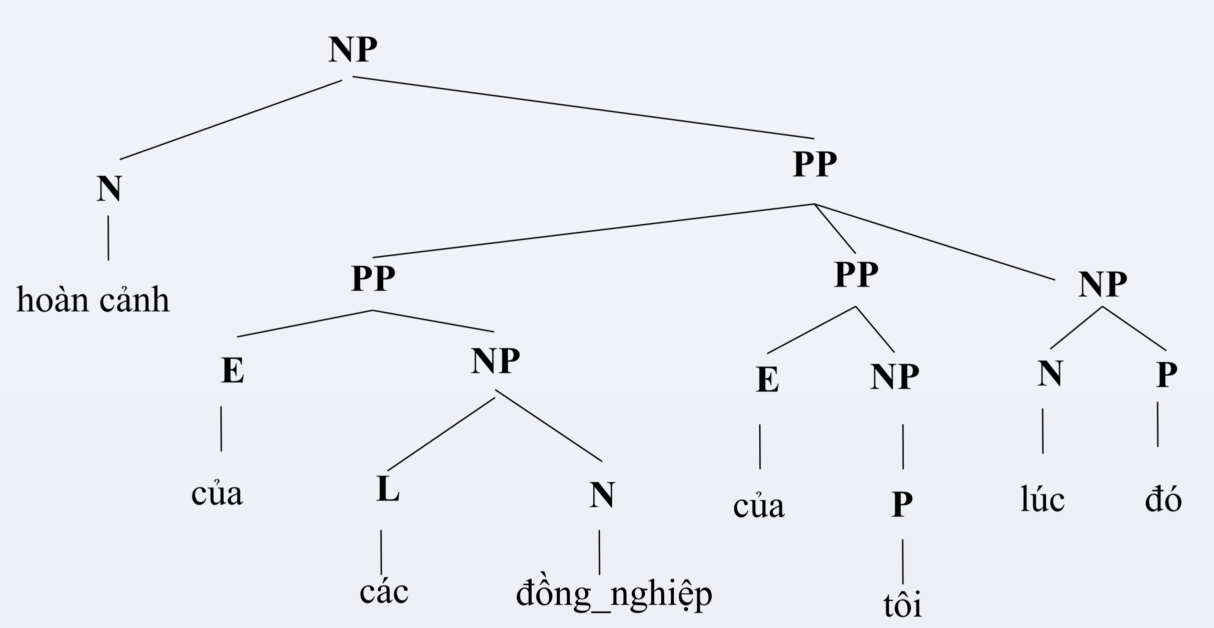
[1]Quangninh University of Industry,
[2]University of Engineering and Technology, Vietnam National Unviversity, Hanoi,
[3]Military Information Technology Institute,
[4]Insitute of Linguistics

Abstract

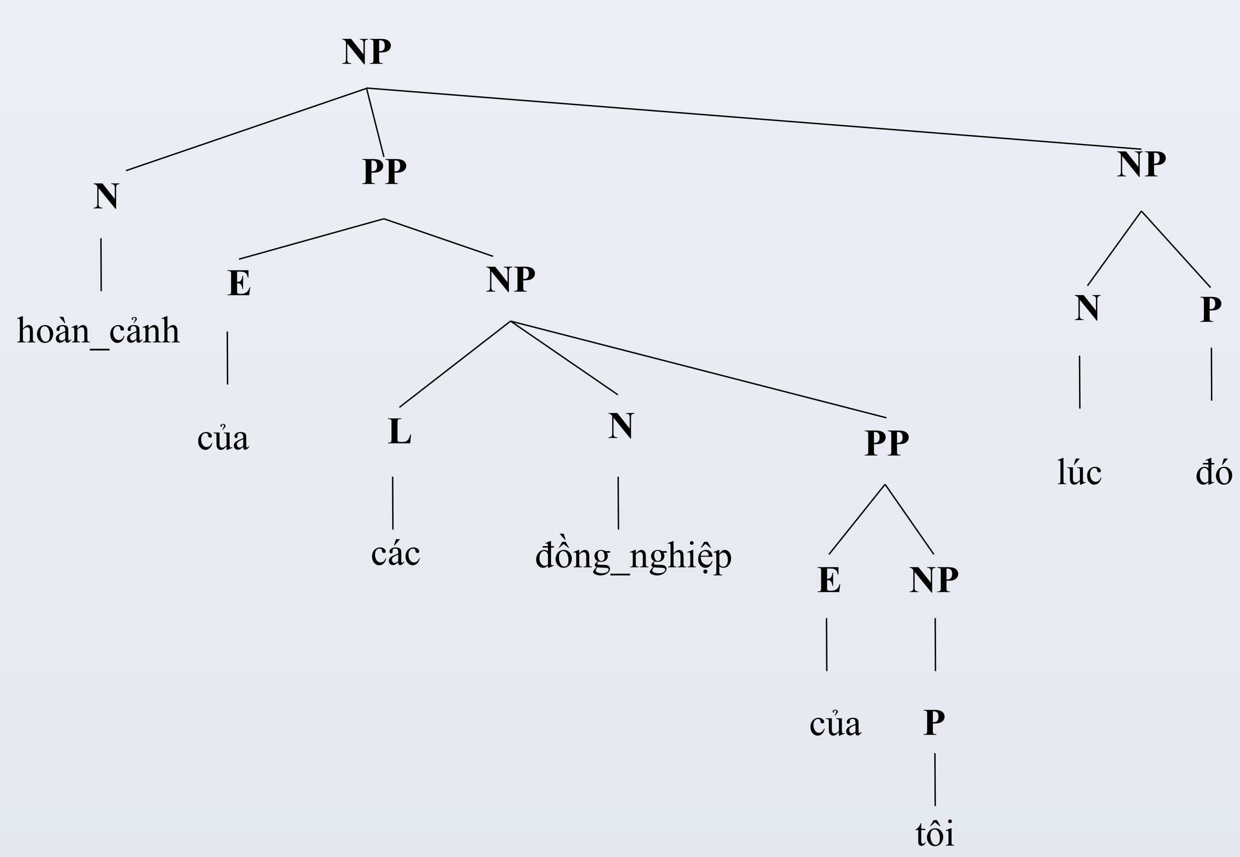
Viet Treebank is an annotated corpus newly published in 2009. In this paper, we applied automated methods to detect errors in Viet Treebank based on the concept of equivalence classes proposed by Dickinson. On this basis, we propose an improved method of error detection by transforming the syntax tree based on vertical markovization. Our experimental results on Viet Treebank showed that the scope of error detection extended more than 2 times and improved the precision more than 18.07% in comparison with the base line methods.

Objectives

The ambiguity in the attached phrase



Correct the ambiguity in the attached phrase



Methods

Proposed a limited concept of equivalence classes to detect rare rules, i.e rules may appear very little in corpus.

-Equivalence Class (EC)

Equivalence classes are conducted according to the following steps:

- Remove daughter categories that are always non-predictive to phrase categorization, i.e., always adjuncts such as punctuation and the parenthetical category.
- Group head-equivalent lexical categories, e.g., N (common noun) and Np (proper noun).
- Model adjacent identical elements as a single element, e.g., NP NP becomes NP.

-Whole daughters scoring (WDS)

- Map the rule to its equivalence class (determine its rule type)
- Score as formula:

$$WDS = |E| + 1/2 \sum_{i=1}^{n} |S_i|$$

When $|E|$ is the number of elements within the equivalence class, $|S_i|$ is the number of elements of the highly similar equivalence class i to the rule type

- Bigram Scoring (BGS)

- Reduce the rule under the concept of limited equivalence class. Resulting in a reduced rule (or a rule type).
- Calculate the frequency of each <mother, bigram> pair in the reduced rule: for search occurrence of <mother, bigram> pair in a class, add a score of 1 for that pair.
- Assign score of the lowest-score <mother, bigram> pair as the score of the rule. We do that because we are interested in anomalous sequences.

Results

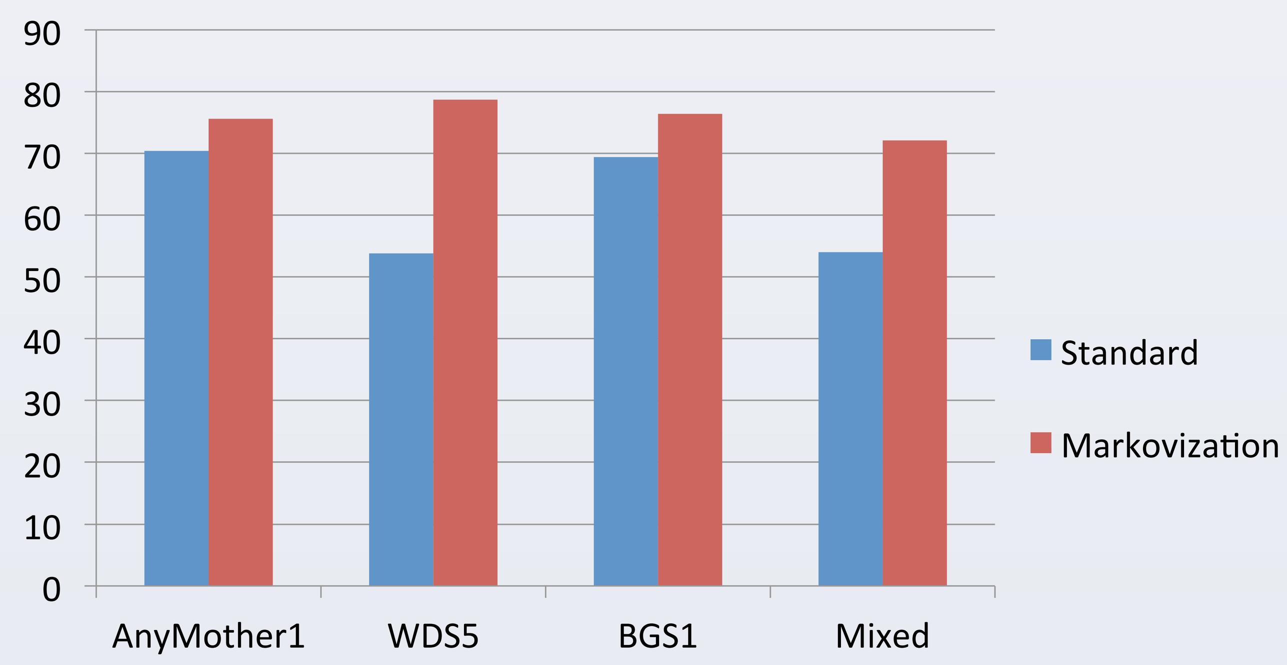
Error detection evaluation

Error set	#Errors detected	#Errors corrected	Precision (%)
EC2	427	298	69,79
EC1	375	264	70,40
WDS10	315	182	57,78
WDS5	157	84	53,50
BGS2	892	597	66,92
BGS1	457	317	69,37
Combined set	1295	700	54,05

Evaluation of error detection with vertical Markovization

Error set	# errors detected	# errors corrected	Precision (%)
EC1	1699	1285	75,63
WDS5	478	376	78,66
BGS1	1516	1158	76,38
Combined set	2791	2013	72,12

Comparison of accuracy before and after Markovization



Conclusion

We found out a significant number of annotation errors in VTB. With these results, we also help linguists that build VTB not only in reducing error detection efforts but also in adjusting annotation guidelines to raise quality of the Vietnamese corpus. The results also showed a significant improvement in error detection scope and precision by transforming syntactic trees to allow additional contextual information about rules to be considered.

References

[1] M.Collins,T. Koo, Discriminative Reranking for Natural Language Parsing, Computational Linguistics 31, 2005.

[2] M. Dickinson, Similarity and dissimilarity in treebank grammars, In Current Issues in Unity and Diversity of Languages: Collection of the papers selected from the 18th International Congress of Linguists (CIL18) , pages 1597–1611, Seoul, 2009.

[3] M. Dickinson, Ad hoc treebank structures, In Proceedings of ACL-08: HLT, pages 362–370, Columbus, Ohio, 2008.

[4] M.Dickinson, Rule equivalence for error detection, In Proceedings of the Fifth Workshop on Treebanks and Linguistic Theories (TLT 2006), pages 187 198. Prague, Czech Republic, 2006.

[5] M.Dickinson and D.Meurers, Prune diseased branches to get healthy trees! How to find erroneous local trees in a treebank and why it matters, Proceedings of TLT 2005, Barcelona, Spain, 2005.

[6] M.Dickinson, and D.Meurers, Detecting Errors in Part-of-Speech Annotation, In Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics, Budapest, Hungary, 2003

[7] M. Dickinson and D. Meurers, Detecting inconsistencies in treebanks, In Proceedings of the Second Workshop on Treebanks and Linguistic Theories (TLT 2003), Sweden. Treebanks and Linguistic Theories, 2003

[8] Guide line of Viet Treebank, VLSP, Project of KC01/06-10, 2010.

[9] R.Jackendoff 1977, X' Syntax: A Study of Phrase Structure. Cambridge, MA: MIT Press.

[10] M.Johnson,PCFG models of linguistic tree representations. ComputationalLinguistics, 24:613-632, 1998

[11] De Kok, J. Ma, &G.VanNoord, A generalized method for iterative error mining in parsing results.Proceedings of the 2009 Workshop on Grammar Engineering Across Frameworks, ACL-IJCNLP 2009, 71-79, 2009

[12] Kvˇeton, Pavel and Karel Oliva, Achieving an Almost Correct PoS-Tagged Corpus, In Text, Speech and Dialogue (TSD), 2002

[13] D.Klein and D. Manning, Accurate Unlexicalized Parsing, In Proceedings of the41st Annual Meeting on Association for Computational Linguistics - Volume 1, ACL '03, pages 423-430, Stroudsburg, PA, USA. Association for Computational Linguistics, 2003

[14] T. Nakagawa , Y Matsumoto, Detecting errors in corpora using support vector machines.In Proceedings of the 19th International Conference on Computatinal Linguistics, pages 709-715, 2002

[15] Gertjan van Noord, Error mining for widecoverage grammar engineering, In Proc. of ACL 2004, Barcelona, Spain, 2004.

[16] Phuong-Thai Nguyen,Anh-Cuong Le, Tu-BaoHo,Thi-Thanh-Tam-Do, Two Entropy-Based Methods for Detecting Errors in POS-Tagged Treebank,. In Proceedings of the Third International Conference on Knowledge and Systems Engineering (KSE), 2011.

[17] Phuong-Thai Nguyen, Xuan-Luong Vu, Thi-Minh-Huyen Nguyen, Van-Hiep Nguyen, Hong-Phuong Le, Building a Large Syntactically-Annotated Corpus of Vietnamese, In Proceedings of LAW-3, ACL-IJCNLP. 2009

www.PosterPresentations.com