Evaluation and Error Analysis of Rule-based Paraphrase Generation for Czech

Veronika Burgerová and Aleš Horák

Natural Language Processing Centre
Faculty of Informatics, Masaryk University
Botanická 68a, 602 00, Brno, Czech republic
{xburger, hales}@fi.muni.cz

Abstract. In this paper, we present the experiments and evaluation of previously developed rule-based paraphrasing system for the Czech language. The system offers several interconnected modules that allow to generate paraphrases of an input sentence based on various criteria such as the Czech WordNet hierarchy, word-ordering rules or anaphora resolution. We have evaluated each module's accuracy and we offer a detailed analysis of the results as well as concrete proposals for improvements.

Keywords: paraphrasing, rule-based, game with a purpose, Czech

1 Introduction

The possibility to programmatically identify or generate paraphrases of an input text allows a plethora of practical natural language processing applications such as machine translation [3], text summarization [15], or semantic interpretation of phrases [6,2]. The techniques for paraphrase generation range from explainable rule-based or thesaurus-based methods [14] to unsupervised approaches usually inspired by machine translation solutions as a monolingual translation [10,9].

In this paper, we evaluate the current results of a previously published rule-base paraphrasing system for Czech [7], which was explicated in a game-with-a-purpose application named Watsonson. We offer a detailed analysis of the results of each of Watsonson's modules and also propose their further development.

2 The Watsonson Project

In general, the task of automatic quality evaluation of a generated sentence is quite difficult. If reference results prepared by human annotators are available, the evaluation can proceed by comparative measures. Without such gold standard datasets, the evaluation mostly relies on human judgement. However, the manual annotation and evaluation of a large set of sentences can be expensive. In the Watsonson project, the paraphrasing results are evaluated in a crowdsourcing approach in the form of a *game with a purpose* (GWAP [1]).

A. Horák, P. Rychlý, A. Rambousek (eds.): Proceedings of Recent Advances in Slavonic Natural Language Processing, RASLAN 2019, pp. 33–39, 2019. © Tribun EU 2019

Table 1: Example paraphrases generated by individual Watsonson modules.

module	sentence
Input	Pejsci si chtěli hrát s **dětmi**, ale žádné děti venku nenašli. (Dogs wanted to play with **children** but they found no children outside.)
wordhyp	Pejsci si chtěli hrát s **lidmi**. (Dogs wanted to play with **people**.)
wordnet	Pejsci si chtěli hrát s **děcky**. (Dogs wanted to play with **kids**.)
wordorder	Nenašli **oni** žádné děti venku. (**They** did not find any children outside.)
aara	**Pejsci** nenašli žádné děti venku. (**Dogs** found no children outside.)
verbinfer	**Neobjevili** žádné děti venku. (They **discovered** no children outside.)

The project uses existing tools such as a morphological analyzer or a syntactic parser and comes up with rule-based procedures to generate paraphrases. The independence of these modules allows us to use and evaluate each module separately.

A detailed description of the Watsonson project in available in [8]. We briefly introduce its five modules we have experimentally evaluated. Example paraphrases generated by the particular modules are presented in Table 1.

2.1 Wordnet and Wordhyp

The *wordnet* module uses data from the Czech WordNet [11] for synonym replacement. Recursively, the words can be replaced by their hypernyms, which is the task of the related the *wordhyp* module. The modules currently do not employ any word sense identification technique to distinguish word senses, which is why wrong paraphrases can be generated.

2.2 Word order

Considering the flexibility of Czech word order, sentence constituents can be reordered in many combinations which still form a correct Czech sentence. This *word order* generates phrases with all possible orders of the sentence constituents.

2.3 Aara

The *Aara* module implements a partial anaphora resolution system. This system resolves zero subjects and replaces pronominal objects or subjects by their co-referent antecedents. In the module, such phrases are generated and offered for annotation.

2.4 Verbinfer

The Czech verb valency lexicon VerbaLex [4] is used in this module, which uses the verb frame inference of three types: equality, effect and precondition to transform the phrases. For instance, *be sad* might be an effect of *get lost*. Besides generating paraphrases, this module can also result in new facts.

Table 2: Statistics of the input testing dataset.

# of sentences	40
# of clauses	97
minimum sentence length	2
maximum sentence length	40
# of words	530
# of pronouns	63
# of named entities	10

3 Experiments and Evaluation

Although the Watsonson users evaluate the sentences given by the paraphrase generator, we can not always distinguish which parts of the system lead to the good evaluations and which ones cause errors. Also, the evaluation of a sentence may be subjective and the decision whether the sentence is correct or not might be ambiguous in some cases.

That is why, the presented evaluation experiment worked with individual modules only and the input paraphrases were processed and by each module separately and tested for correctness.

3.1 Preprocessing

The first phase of the experiment consisted in taking 40 Czech sentences of various complexity. The sentences were either simple made up phrases or they were extracted from Czech children tales. Detailed statistics of this dataset are displayed in Table 2.

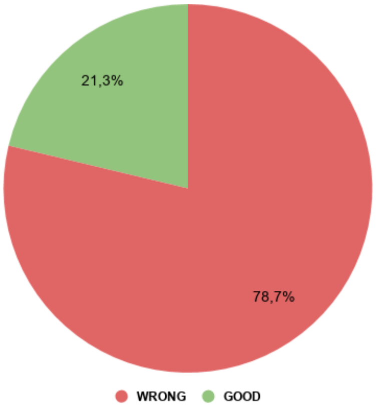

Fig. 1: The total score of the five modules.

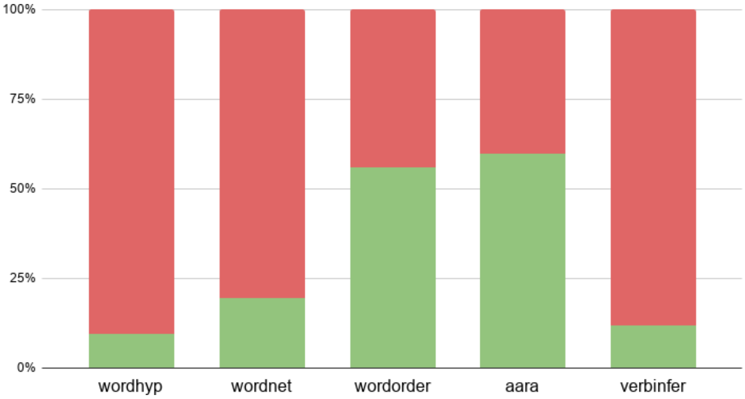

Fig. 2: The individual scores of each module.

As the individual modules rely on morphological and syntactic annotations, the sentences we first processed by the annotating tools: morphological analyzer/generator *majka* [12], morphological tagger *desamb* [13] and the syntactic parser *SET* [5].

The preprocessing pipeline results with tagged and parsed sentences were stored in the JSON format and server as an input to each of the evaluated modules. In total, the paraphrasing modules generated 1,514 new sentences based on the 40 original statements.

3.2 Evaluation

Within the evaluation, all the generated paraphrases were manually annotated. At first, each sentence was marked whether it is or is not a good paraphrase of its input.

As can be seen in Figure 1, only 21.3 % of the 1514 paraphrases were evaluated as good paraphrases. Comparing this score with the score of the paraphrases

Table 3: The statistics of the paraphrasing results per module.

module	# of good paraphrases		# of bad paraphrases		total # of paraphrases	
wordhyp	52	← 9%	505	←91%	557	↓37%
wordnet	95	←19%	393	←81%	488	↓32%
wordorder	93	←56%	73	←44%	166	↓11%
aara	58	←60%	39	←40%	97	↓ 6%
verbinfer	24	←12%	182	←88%	206	↓14%
total	322	21%	1,192	79%	1,514	100%

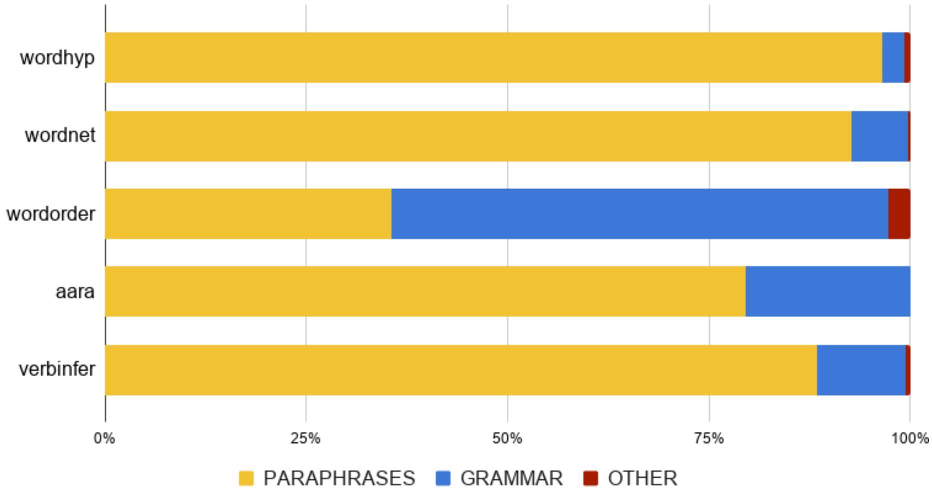

Fig. 3: The results of the error analysis applied on each of the modules.

evaluated by Watsonson users, which is more than 55 %, the total score is lower than expected.

Focusing on the modules' results separately in Figure 2, we are able to see which modules are beneficial to the paraphrase generation process and which ones take the total score down. The score ratios and numbers of sentences generated by the particular modules are presented in Table 3.

3.3 Error analysis

After the basic evaluation, the next step focussed on detailed analysis of the errors in the generation process. The errors were classified in the following three categories:

- **Incorrect paraphrase**. This means that the generated sentence does not make sense at all or does not follow the meaning of the original sentence.
- **Incorrect grammar**. The sentence meaning is synonymous to the meaning of the original sentence and it would make a good paraphrase, but it is not grammatically correct.
- **Other**. Unspecified type of an error not fitting any of the types mentioned above.

The results of this detailed error analysis for each module are presented in Figure 3. Most of the errors are caused by incorrect paraphrasing, however, we can observe that a significant percentage of errors is grammatical, especially in the *wordorder* module.

In the last part of the analysis, we focused on the most common errors that occurred in the results and tried to find out the source of such errors.

As we have shown, most of the incorrect paraphrases were generated by the *wordhyp*, *wordnet* and *verbinfer* modules. All of these modules generate paraphrases on the basis of replacing words in a sentence by words with similar meaning. Nevertheless, these meanings often do not fit in the given context.

On the contrary, most of the errors made by the *wordorder* module, were caused by forming an ungrammatical sentence. Nevertheless, the overall score of this module is comparatively high due to the fact that the Czech word order is very flexible, but it is not completely free.

After the individual evaluation of each module separately, we have analysed the errors that occurred repeatedly across all the modules and identified the main reasons of them:

1. **Prepositions**. We have noticed several paraphrases that were either missing a preposition where it was needed or occurred with incorrect or redundant preposition.
2. **Dependencies**. In a lot of paraphrases, we observed incorrect dependencies among the parts of a sentence, for instance, object generated as subject etc.

Further analysis of the primary source if these errors revealed that a significant part of these errors was being caused by errors in the syntactic parsing phase.

4 Conclusions and Future Work

We have presented a detailed evaluation and error analysis of five paraphrase generation modules of the Watsonson project. The analysis showed the most problematic sources of errors in the generation process and helped to pave road for further improvements of the system.

New modules are planned to provide new types of paraphrasing methods such as replacing other sentence constituents than nouns and verbs or transforming the active voice to passive and vice versa.

Acknowledgements This work has been partly supported by the Ministry of Education of CR within the LINDAT-Clarin infrastructure LM2015071.

References

1. von Ahn, L.: Games with a purpose. Computer **39**(6), 92–94 (2006)
2. Bollegala, D., Shutova, E.: Metaphor interpretation using paraphrases extracted from the web. PloS one **8**(9), e74304 (2013)
3. Callison-Burch, C., Koehn, P., Osborne, M.: Improved statistical machine translation using paraphrases. In: Proceedings of the main conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics. pp. 17–24. Association for Computational Linguistics (2006)

4. Hlaváčková, D., Horák, A., Kadlec, V.: Exploitation of the VerbaLex verb valency lexicon in the syntactic analysis of Czech. In: International Conference on Text, Speech and Dialogue, TSD 2006. pp. 79–85. Springer (2006)
5. Jakubíček, M., Horák, A., Kovář, V.: Mining phrases from syntactic analysis. In: International Conference on Text, Speech and Dialogue, TSD 2009. pp. 124–130. Springer (2009)
6. Nakov, P.I., Hearst, M.A.: Semantic interpretation of noun compounds using verbal and other paraphrases. ACM Transactions on Speech and Language Processing (TSLP) **10**(3), 13 (2013)
7. Nevěřilová, Z.: Paraphrase and textual entailment generation. In: International Conference on Text, Speech, and Dialogue, TSD 2014. pp. 293–300. Springer (2014)
8. Nevěřilová, Z.: Annotation game for textual entailment evaluation. In: Gelbukh, A.F. (ed.) Proceedings of CICLing. LNCS, vol. 8403, pp. 340–350. Springer, Heidelberg (2014)
9. Prakash, A., Hasan, S.A., Lee, K., Datla, V., Qadir, A., Liu, J., Farri, O.: Neural paraphrase generation with stacked residual lstm networks. arXiv preprint arXiv:1610.03098 (2016)
10. Quirk, C., Brockett, C., Dolan, W.: Monolingual machine translation for paraphrase generation. In: Proceedings of the 2004 conference on empirical methods in natural language processing. pp. 142–149 (2004)
11. Rambousek, A., Pala, K., Tukačová, S.: Overview and Future of Czech Wordnet. In: McCrae, J.P., Bond, F., Buitelaar, P., Cimiano, P., 4, T.D., Gracia, J., Kernerman, I., Ponsoda, E.M., Ordan, N., Piasecki, M. (eds.) LDK Workshops: OntoLex, TIAD and Challenges for Wordnets. pp. 146–151. CEUR-WS.org, Galway, Ireland (2017), `http://ceur-ws.org/Vol-1899/`
12. Šmerk, P.: Fast Morphological Analysis of Czech. In: Proceedings of Recent Advances in Slavonic Natural Language Processing, RASLAN 2009. pp. 13–16 (2009)
13. Šmerk, P.: K počítačové morfologické analýze češtiny (in Czech, Towards Computational Morphological Analysis of Czech). Ph.D. thesis, Faculty of Informatics, Masaryk University (2010)
14. Zhao, S., Lan, X., Liu, T., Li, S.: Application-driven statistical paraphrase generation. In: Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2-Volume 2. pp. 834–842. Association for Computational Linguistics (2009)
15. Zhou, L., Lin, C.Y., Munteanu, D.S., Hovy, E.: Paraeval: Using paraphrases to evaluate summaries automatically. In: Proceedings of the main conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics. pp. 447–454. Association for Computational Linguistics (2006)