Can Large Language Models Tell Us Something about Derivation Processes?

Marko Tadić
University of Zagreb,
Faculty of Humanities and Social Sciences,
Zagreb, Croatia
marko.tadic@ffzg.unizg.hr

Abstract

The paper presents the preliminary research on usage of Large Language Models (LLMs), primarily using translation models in Neural Machine Translation (NMT) process, to generate newly derived and compound words. The method for detecting and classifying newly generated words by usage of NMT translation models, is being presented.

1 Introduction

Recently there has been a clear shift from knowledge-based and human-engineered methods towards data-driven architectures, which has led to the progress in the field of Language Technology (LT). One recent aspect associated with the paradigm shift in language processing is the use of pretrained Large Language Models (LLMs). Large-scale monolingual and/or multilingual textual data is used to train LLMs. Pre-trained LLMs, like BERT (Devlin et al., 2019), GPT-3 (Brown et al., 2020), GPT-4 (OpenAI, 2023), and XLM-RoBERTa (Conneau et al., 2020), have offered a framework for using the knowledge acquired during the training process to be later applied to newer tasks. One such task could be the usage of LLMs in detection of derivational morphology phenomena and, if possible, their classification and description. In this respect this paper presents results of a preliminary research that tries to determine whether a LLM can be used to detect derivationally and compositionally newly generated words in a language. The detection process in essence boils down to the usage of a Neural Machine Translation (NMT) model pretrained on parallel data (Croatian-English parallel corpus) to translate one side of already humanly translated Croatian-English parallel corpus: English into Croatian again. The resulting translation has been matched with the existing Croatian lexica in order to detect newly coined words, i.e. words that are unknown to the existing lexical resources. These words are distributed in several categories, their overall frequency is presented and the results are being discussed.

The paper is structured as follows. The section 2 presents the scarce related work where LLMs were used in derivational morphology, while in the section 3 the used language resources are described. In the section 4 the methodology is detailed and in section 5 results are presented accompanied by discussion. The conclusions and possible future directions of research are provided in section 6.

2 Related Work

So far the usage of LLMs in processing derivational morphology has been quite scarce. Cotterell et al. (2017) and Deutsch et al. (2018) proposed neural architectures that represent derivational meanings as tags. In Edmiston (2020) experiments, which probe the hidden representations of several BERT-style models for morphological content, are being presented and discussed. The most prominent work in this task so far is provided by Hofmann et al. (2020a, 2020b, 2021) where the authors use the auto-encoder to check the morphological well-formedness (MWF), finetune the BERT into DagoBERT that is capable of generating new derivations, and use finetuning to improve

Proceedings of the 4th Int. Workshop on Resources and Tools for Derivational Morphology (DeriMo 2023), pages 33–38,
Dubrovnik, Croatia, 5-6 October 2023.

BERT's interpretation of complex words.

All these papers based their research on monolingual LLMs, while to the best of our knowledge, the approach proposed in this paper is the first one that uses the LLMs in multilingual context, i.e. using the translation model and LLM for the target language in order to investigate and extrinsically evaluate the generation of derivatives and compounds in a target language. Starting from the parallel corpus enables us to keep the content variable under control since in a parallel corpus the translation equivalents at the level of sentences, i.e. translation units (TUs), are explicitly marked by unique sentence IDs and are considered to convey the same overall sentence meaning.

3 Language Resources

In this research we used the following language resources:

The parallel corpus that was used for experiments is the Croatian-English Parallel Corpus (CW) described in (Tadić, 2000). It is a unidirectional corpus of newspaper articles published in *Croatia Weekly* between 1998 and 2000, translated by professional translators and language proofed by three different English native speakers.

For the machine translation of English sentences into Croatian, the NMT models developed within the CEF-project National Language Technology Platform (NLTP)[1] were used through the online interface[2] of its Croatian installation (Vasiļevskis et al. 2023). The baseline NMT model were trained on DGT parallel English-Croatian data and then finetuned with additional 0.76 million Translation Units (TUs), including the whole CW parallel data set. The NMT models are typical Transformer based models that were produced by Tilde[3] are described in (Krišlauks and Pinnis 2020). However, unlike in the described en→pl translation model training, that also used backtranslation due to the noisy input data, for training en→hr and hr→en translation models only the Transformer base model configuration was used since the data were composed of only clean human translations. The first instance of these translation models for en→hr and hr→en pairs was used in EU Council Presidency Translator[4] in 2020 and it received BLEU scores of 36.93 and 41.30 respectively. For the NLTP Croatian installation, these translation models were enriched with additional training data of approximately 1 million tokens and has been used in this experiment.

For tokenization of translated sentences the UDPipe pipeline[5] has been used with the Croatian set UD2.10 selected.

For detection of the unknown words the Croatian Morphological Lexicon (HML)[6], an inflectional lexicon with 110,000+ lemmas and 6M+ wordforms, accessible as an online service, was used. Its features were described in detail in (Tadić, 2005).

During the checking of unknown words a number of existing Croatian language resources were used starting with corpora: Croatian National Corpus (HNK)[7] and Croatian Web Corpus (hrWaC)[8]. The online lexica used for checking were: Hrvatski jezični portal[9], Croatian Special Field Terminology[10], Croatian Terminology Portal[11], Jezikoslovac[12], Croatian Glosbe[13] online dictionary, CroDict[14] online dictionary, Croatian Encyclopedia[15], co-textual search engine Kontekst[16] set to Croatian queries, and common search engines Google and DuckDuckGo. Also, as the final means

[1] https://nltp-project.info
[2] https://hrvojka.gov.hr
[3] https://www.tilde.com
[4] https://hr.presidencymt.eu
[5] https://lindat.mff.cuni.cz/services/udpipe/
[6] https://hml.ffzg.hr
[7] https://filip.ffzg.hr
[8] http://nlp.ffzg.hr/resources/corpora/hrwac/
[9] https://hjp.srce.hr
[10] https://struna.ihjj.hr
[11] https://nazivlje.hr
[12] https://jezikoslovac.com
[13] https://hr.glosbe.com
[14] https://crodict.hr
[15] https://enciklopedija.hr
[16] https://kontekst.io

used for finding a lexical evidence the paper version of the *Veliki rječnik hrvatskoga standardnoga jezika* (Jojić et al., 2015) was used.

4 Methodology

In this section the methodology used in the research is described in detail.

4.1 Translation of English Sentences

The CW was obtained from META-SHARE[17] in TMX format and the sample of 10,000 TUs was selected and English sentences from aligned pairs were extracted. The unique sentence IDs were preserved in order to be able to refer back to the original Croatian sentences (*hr* mark in examples) when needed.

The English sentences were translated using the Croatian installation of the NLTP NMT services at hrvojka.gov.hr. The source 10,000 English TUs (*en* mark in examples) had 234,278 tokens, while the translated Croatian TUs (*hr-t* mark in examples) had 193,020 tokens.

4.2 Tokenisation with UDPipe and Matching with the Croatian Morphological Lexicon

The *hr-t* set of sentences was tokenized using the UDPipe online services and the results were downloaded in CoNLL-U format. Only the first column of that format was used in this research. However, the annotation information from the remaining nine columns could be used for future investigations on e.g. quality of lemmatization, particularly when it comes to the unknown and for UDPipe system unseen words. This might be one of directions for the future research, but it certainly surpasses the limits of the current paper.

The token list from the first column was uploaded to the HML requesting the lemmatization of each token. In the case of unknown token, the HML returns #NIL#, so it was easy to extract words unknown to HML.

4.3 Inspection and Classification of Unknown Words

The list of #NIL# tagged tokens, 4453 in total, was then manually inspected for evidence. Every token not being evidenced in any of aforementioned corpora, lexica or search engines was marked and classified in accordance with the preliminary classification scheme. The scheme and basic statistics is presented in Table 1.

Before the manual inspection it was decided that certain types of unknown words will not be taken into account: 1) named entities; 2) translation errors (e.g. direct transfer of the original English word); 3) deverbative nouns ending in *–nje* since they are highly productive in Croatian[18]; 4) highly productive negated adjectives and nouns (e.g. *nekoristan*, *nekompetencija*); 5) highly productive compounds written usually with dash (e.g. *makedonsko-hrvatski*, *ne-Hrvat*). On the other hand, we put a strong emphasis on detecting compounds written without dash since they usually express stronger bond between compounding parts.

5 Results and Discussion

Here each of the categories of the classification scheme is described and exemplified. :

- expectable compound: compounds that could be expected having in mind possible combination of compounding parts, e.g. en: *self-denying* / hr-t: *samoopovrgavajući*, en: *late antique* / hr-t: *kasnoantika*
- unexpectable compound: compounds that are partial errors in translation but convey the general meaning, e.g. en: *five-movement* / hr-t: *petokretni* instead of hr: *petostavačni*, en: *Euro game* / hr-t: *euroigre* instead of hr: *europske igre*;

[17] https://meta-share.org

[18] This decision could be questioned since investigating this highly regular and productive derivational pattern in Croatian (and many other Slavic languages as well) could reveal some of the underlying mechanisms that LLMs are dealing with when trained at the subword level. However, this topic might deserve the paper on its own while here we wanted to tackle the widest possible number of different phenomena at this preliminary pilot level.

- possessive adjective of names: highly productive derivation, but sometimes with unexpected derivations, e.g. en: *Boka Croats* / hr-t: *bočki Hrvati* instead of hr: *Hrvati iz Boke*, en: *Klein's* / hr-t: *Kleinski* instead of hr: *Kleinov*;
- alternative derivation: derivation that uses different, but possible, derivation affix, e.g. en: *lace-makers* / hr-t: *čipkaši*, en: *broker* / hr-t: *burzer*;
- unexpectable derivation: derivations that are partial errors in translation, but convey the general or alternative meaning, e.g. en: *swallow* (bird) / hr-t: *gutljica*, en: *voucher holders* / hr-t: *imatelji vaučera*;
- direct alternative calque: derivations or compounds that directly conveys the English word and tries to translate its parts and/or adapt it phonetically and morphologically in Croatian, e.g. en: *underworld organisations* / hr-t: *podsvjetske organizacije* instead of hr: *mafijaške organizacije*, en: *Knights Hospitallers* / hr-t: *Hospitalari* instead of hr: *ivanovci*.

Category	Tag	Frequency
expectable compound	so	17
unexpectable compound	sn	11
possessive adjective (-ov/-ski/-čki)	pp	164
alternative derivation	dz	76
unexpectable derivation	dn	15
direct alternative calque	pz	38
total		321

Table 1: Words unknown to the existing lexica and their classification scheme with basic statistics.

The initial 4453 words marked with #NIL# as the result of matching with HML, were scaled down after the manual inspection and lookup for evidence in different language resources to the total of 321 cases. Most of the tokens unknown to HML were named entities and clear translation errors.

The 321 occurrences of newly generated words represent 7,21% of all unknown words. This might look like a small number, but this should be regarded as a percentage of total number of lexical entries used in the sampled 10,000 TUs, i.e. 193,020 hr-t tokens. These 7,21% cases are the spots in the English text that for some reason invoked the translation LM to come up with derivation or composition in order to convey the basic meaning. Was it invoked because of the lacunae in Croatian lexicon where in the English such lexical items exist? Certainly not since the manual inspection confirmed that in many cases in the original Croatian source such lexical items exist.

Does the LLMs have intrinsic preference to generate derivations or compounds because of limited lexicon used in the training process? What is really being conveyed with this language means and their selection in the process of machine translation using LLMs? Is it the similar content running in two parallel texts, or approximation of its similarity represented through LLM-based MT, that affects also such lower language levels as derivational morphology?

The individual examples for expectable categories might look quite surprising to a native speaker of Croatian, but after careful inspection of the English source, the expected and alternative derivatives and compounds generated in translations are morphologically well-formed (see examples above).

6 Conclusions and Future Directions

We presented the preliminary investigation that tried to detect the amount and types of possible newly derived and compound words produced by a LLM. The LLMs (particularly translation

models where we have the experimental variable of the same content in two languages under control), that are being trained to take into account the subword segments, in their performance are now being able to signal the spots where the transferred content could be represented by derivational or compositional means available in the target language. In this respect the LLM generates the derivations or compositions not yet registered in any lexica of the target language by following the derivational and compositional rules of that language and thus producing MWF words. At this spots are LLMs signaling us something? It seem like they are pinpointing the nodes in the total combinatorial capacity of a language at the derivational/compositional level, the nodes in the derivational/compositional network of morpheme combinations, that exist *in potentia*, but are not (yet) filled with an accepted combination of morphemes. These nodes were certainly not filled with lexical entries in the training material, but still the LLM has envisaged their existence. Can LLMs help us in recognizing the topology of this network or it is just another way of representation of the derivational/compositional complexity in language?

This production of neologisms is particularly characteristic for translation pair en→hr since these two languages differ typologically, namely English is more analytical and tends towards phrasal solutions, while Croatian is more synthetical and tends towards derivational solutions. It would be interesting in the future to investigate the reverse direction of translation, i.e. hr→en and then check the ability of the same translation LM to generate derivatives/compounds in English and to provide their classification and statistics.

If humans would generate such new words, representing in fact new lexical entries, we would tend to consider this a creative use of language. Can we treat such words the same way when they are being generated by LLMs?

Although the research presented in this paper didn't produce fully automatic method of detecting newly generated derivations and compositions, this could represent one of directions for future research. We have a parallel corpus at our disposal and the difference between the humanly produced original text in Croatian and NMT produced translated counterpart from English into Croatian could be automatically compared for differences.

Moreover, following (Hofmann et al. 2020b), we need further intrinsical evaluation to find out how input segmentation impacts the derivational knowledge available to a LLM. This might suggest that the performance of LLMs could be improved if a morphologically informed vocabulary of units (e.g. derivationally segmented) were used in the training phase. At this stage of training of LLMs, we don't really know how the subword segmentation is being produced and to what extent the division into segments really corresponds to the real morphological boundaries.

It would certainly be most useful if we could make use of existing LLMs in the computational processing of derivational/compositional morphology and even more so if we could perhaps be able to train a new LLM tailored to be sensitive on derivational/compositional information.

Acknowledgments

The work reported here was supported by the European Commission through the CEF Telecom Programme (Action No: 2020-EU-IA-0082 National Language Technology Platform, NLTP, Grant Agreement No: INEA/CEF/ICT/A2020/2278398) and by the Ministry of Science and Education of the Republic of Croatia through the support for the Croatian CLARIN Research Infrastructure Consortium.

References

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D. Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, Dario Amodei. 2020. Language models are few-shot learners. Advances in neural information processing systems, 33:1877–1901. https://doi.org/10.48550/arXiv.2005.14165

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco

Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised Cross-lingual Representation Learning at Scale. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 8440–8451, Online. Association for Computational Linguistics. [https://aclanthology.org/2020.acl-main.747.pdf]

Ryan Cotterell, Ekaterina Vylomova, Huda Khayrallah, Christo Kirov, and David Yarowsky. 2017. Paradigm completion for derivational morphology. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 714–720, Copenhagen, Denmark, Association for Computational Linguistics. [https://aclanthology.org/D17-1074.pdf]

Daniel Deutsch, John Hewitt, and Dan Roth. 2018. A distributional and orthographic aggregation model for English derivational morphology. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1938–1947, Melbourne, Australia. Association for Computational Linguistics. [https://aclanthology.org/P18-1180.pdf]

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pages 4171–4186, Minneapolis, Minnesota, Association for Computational Linguistics. [https://aclanthology.org/N19-1423.pdf]

Daniel Edmiston. 2020. A Systematic Analysis of Morphological Content in BERT Models for Multiple Languages. https://doi.org/10.48550/arXiv.2004.03032.

Valentin Hofmann, Hinrich Schütze, and Janet Pierrehumbert. 2020a. A Graph Auto-encoder Model of Derivational Morphology. *In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1127–1138, Online. Association for Computational Linguistics. [https://aclanthology.org/2020.acl-main.106.pdf]

Valentin Hofmann, Janet Pierrehumbert, and Hinrich Schütze. 2020b. DagoBERT: Generating Derivational Morphology with a Pretrained Language Model. *In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3848–3861, Online. Association for Computational Linguistics. [https://aclanthology.org/2020.emnlp-main.316.pdf]

Valentin Hofmann, Janet Pierrehumbert, and Hinrich Schütze. 2021. Superbizarre Is Not Superb: Derivational Morphology Improves BERT's Interpretation of Complex Words. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3594–3608, Online. Association for Computational Linguistics. [https://aclanthology.org/2021.acl-long.279.pdf]

Ljiljana Jojić, Nada Vajs Vinja, Vesna Zečević, Anuška Nakić, Ivan Ott, Jelena Cvitanušić Tvico, Ranka Đurđević, Igor Marko Gligorić, Aida Korajac, Ines Kotarac, Ivana Krajačić, Ivan Ott, Katja Peruško, Nika Štriga, Dijana Vlatković. 2015. *Veliki rječnik hrvatskoga standardnoga jezika*. Zagreb: Školska knjiga.

Rihards Krišlauks, and Mārcis Pinnis. 2020. Tilde at WMT 2020: News Task Systems. In *Proceedings of the 5th Conference on Machine Translation (WMT)*, pages 175–180, Online. Association for Computational Linguistics. [https://aclanthology.org/2020.wmt-1.15.pdf]

Marko Tadić. 2000. Building the Croatian-English Parallel Corpus. In Proceedings of the Second International Conference on Language Resources and Evaluation (LREC'00), pages 523-530, Athens, Greece. European Language Resources Association (ELRA). [http://www.lrec-conf.org/proceedings/lrec2000/pdf/119.pdf]

Marko Tadić. 2005. The Croatian Lemmatization Server. *Southern Journal of Linguistics*, 29(1/2): 206-217.

Artūrs Vasiļevskis, Jānis Ziediņš, Marko Tadić, Željka Motika, Mark Fishel, Bjarni Barkarson, Claudia Borg, Keith Aquilina, and Donatienne Spiteri. 2023. National Language Technology Platform (NLTP): The Final Stage. In *Proceedings of the International Conference HiT-IT2023*, pages 203-208, Naples, Italy. INCOMA Ltd., Šoumen, Bulgaria. doi:10.26615/issn.2683-0078.2023_019.