Can Large Language Models Tell Us Something about Derivation Processes?

Marko Tadić

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

DeriMo2023, Dubrovnik, 2023-10-05

Overview

- introduction
- Language Resources (LRs) used
- methodology
- results & discussion
- conclusions

Introduction

- Large Language Models (LLMs)
  - large-scale monolingual and/or multilingual textual data used to train LLMs
  - pre-trained LLMs use the “knowledge” acquired during the training to be applied to new tasks
- can LLMs be used in
  - detection of derivational morphology phenomena
  - their classification
  - their description?
- results of a preliminary research
  - detecting whether LLM can generate derivationally and compositionally new words in a language

Introduction 2

- related works
  - so far used only monolingual LLMs
- here we have a multilingual environment
  - NMT translation model and connected LLM for smoothing the target language output
  - to investigate and extrinsically evaluate the generation of derivatives and compounds in a target language
  - translation LMs are geared towards the generation
  - parallel corpus allows us to have a content variable under control
- motivation
  - back in 2020: a case of en: *three goal* NMT translated to hr *trogol*

Language Resources

- Croatian-English Parallel Corpus
  - Tadić (2000): 3.5 Mw unidirectional parallel corpus of *Croatia Weekly* (CW) newspapers, 1998-2000, hr original translated into en by professional translators, en proofed by three native speakers
  - ***hr***: original; ***en***: human translation to en; ***hr-t***: NMT translation to hr
- Hrvojka: NMT translation system (hrvojka.gov.hr)
  - Vasiļevskis et al. (2023): result of the CEF-project NLTP
  - Krišlauks & Pinnis (2020): NMT models
    - typical Transformer based models
    - produced by Tilde, won the 2017-2019 WMT competition
    - no backtranslation needed: HQ human translation without noise

Language Resources 2

- UDPipe for tokenisation
  - Straka & Straková (2022): Croatian set UD2.12 selected
- Croatian Morphological Lexicon (HML)
  - Tadić (2005): HML v5, an inflectional lexicon with more than 6 milion entries, i.e. generated word-forms of 110,000+ lemmas
  - used for matching the tokens with the known words
- Croatian corpora
  - Croatian National Corpus (HNK), hrWaC, Riznica
- Croatian lexica
  - most used on/offline lexica: HJP, StruNa, Nazivlje, Jezikoslovac, ...

Methodology

- translation (NMT):
  - en→hr: en part of aligned sentences in CW
  - set of 10,000 sentences
  - en: 234,278 tokens, hr-t: 193,020 tokens
- tokenisation of hr-t:
  - UDPipe online version: http://lindat.mff.cuni.cz/services/udpipe/
- matching with HML
  - token list uploaded to HML: http://hml.ffzg.hr
  - 4453 types unknown to HML, i.e. marked with #NIL#

Methodology 2

- before manual inspection: criteria for exclusion
  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;
  4. highly productive negated adjectives and nouns (e.g. *nekoristan*, *nekompetencija*);
  5. highly productive compounds written usually with dash (e.g. *talijansko-hrvatski*, *ne-Hrvat*).
- manual inspection:
  - most of 4453 types = translation errors or NEs unknown to HML

Results

- after manual inspection
  - of detected 4453 types unknown to HML
    - translation errors or NEs were excluded
    - types confirmed in other lexica, but unknown to HML were excluded
  - 4453 types scaled down to 321 words
  - 321 words
    - 7.21% of all 4453 “unknown” types
    - only 0.166% of all 193,020 hr-t tokens
  - no repetition of generated words observed
    - all occurrences were hapax legomena
    - no multiple word-forms or more than one occurence of the same lemma

Results 2: preliminary classification scheme

1. **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*
2. **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 games* / hr-t: *euroigre* instead of hr: *europske igre*;
3. **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* or *bokeljski Hrvati*, en: *Klein's* / hr-t: *Kleinski* instead of hr: *Kleinov*;

Results 2: preliminary classification scheme 2

4. alternative derivation

- derivation that uses different, but possible, derivation affix
- e.g. en: *lace-makers* / hr-t: *čipkaši*, en: *broker* / hr-t: *burzer*;

5. unexpectable derivation

- partial errors in translation, but convey the general/alternative meaning
- e.g. en: *swallow* (bird) / hr-t: *gutljica*, en: *(voucher) holders* / hr-t: *imatelji (vaučera)*;

6. direct alternative calque

- derivation/compound that directly conveys the English word or 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*.

Results 3

- statistics

Category	Tag	Frequency
1. expectable compound	so	16
2. inexpectable compound	sn	15
3. possessive adjective (*-ov*/*-ski*/*-čki*)	pp	164
4. alternative derivation	dz	65
5. inexpectable derivation	dn	23
6. direct alternative calque	pz	38
Total		**321**

Discussion

- 321 words
  - mark the spots in the English text that induced the translation LM to come up with derivation or composition in order to convey the meaning from en into hr-t
  - following the derivational/compositional rules of the target language producing MWF words
- new derivative/compound generated
  - because of the lacunae in Croatian lexicon while in English lexicon such lexical items exist?
  - not really: manual inspection confirmed that in most cases in the original Croatian source such lexical items exist
  - LM was motivated to generate new word for some other reason

Discussion 2

- is LLM generating new words actually signalling the nodes?
  - in the derivational/compositional network
  - representing the total combinatorial capacity of a language at derivational/compositional level
  - i.e. morpheme combinations that exist *in potentia*
  - Halle (1973):
    - “list of morphemes together with the rules of formation define the set of potential words of a language”
  - combination of units following rules of combination can also be represented as a network, i.e. static vs. processual representation
- Can LLMs help us in recognizing the topology of this network?

Future directions

- investigate highly productive deverbative nouns ending in *-nje* to check this derivational pattern
- reverse the direction of translation: hr→en NMT translation
  - en more analytical, more phrasal solutions vs. hr more synthetical, more derivational solutions
  - check the ability of the same translation LM to generate derivatives/compounds in en
- automatic method of comparison of hr and ht-t
  - parallel corpus will allow us to compare automatically hr and hr-t

Future directions 2

- intrinsical evaluation of LLMs
  - how input segmentation during the training phase impacts the derivational “knowledge” available to a LLM
  - at this stage we don't really know how the subword segmentation is being organised during the training phase of LLMs
  - to what extent the division into segments really corresponds to the real morpheme boundaries
  - we need means to evaluate the performance of LLMs, i.e. benchmarks for carefully tailored prompting

Conclusions

- performance of LLMs, particularly NMT systems, could be improved by additional fine-tuning the LM with in-domain terminological sets
- Can a “morphologically informed” vocabulary (e.g. derivationally segmented) be used in the training phase to fine-tune the LM for derivational morphology processing?
- we could perhaps train a new LLMs tailored to be sensitive on derivational/compositional information

Conclusions 2

- humans generating new words, i.e. new lexical entries
  - we consider this a creative use of language to a certain extent
- Can we treat such words as creative usage of language when they are being generated by LLMs?

Thank you for your attention.

This research has been supported by the:

EU CEF Telecom Programme, Action NLTP

Ministry of Science and Education of the Republic of Croatia
to the HR-CLARIN consortium.