Analyzing the Understanding of Morphologically Complex Words in Large Language Models

Marion Di Marco[1,2] and Alexander Fraser[1,2,3]

[1]School of Computation, Information and Technology, TUM
[2]Center for Information and Language Processing, LMU Munich
[3]Munich Center for Machine Learning, Germany
{marion.dimarco,alexander.fraser}@tum.de

Introduction and Motivation

- Complex morphology is challenging for NLP: much information condensed into one word
- Productive word formation processes → novel or infrequent words
- Ability to model new words: important for morphologically rich languages
- LLMs encode information about word relations – what about word-internal structure?

→ **To what extent can LLMs understand the structure of complex words?**
Design tasks to study compositional word formation and derivation in German

Morphology

- Language-specific patterns to create word forms
- **Derivation:** create word forms through prefixes/suffixes: *(to) use → usable → usability*
- **Compounding:** concatenation of existing words to form new words

Nutzbarkeit + Dauer → *Nutzbarkeitsdauer* (*usability time*)
Gesamt + Nutzbarkeit + Dauer → *Gesamtnutzbarkeitsdauer* (*total usability time*)
Gesamt + Nutzbarkeit → *Gesamtnutzbarkeit* (*total usability*)
Gesamt + Nutzbarkeit + Dauer + Sudie → *Gesamtnutzbarkeitsdauerstudie* (*total usability time study*)

Dataset and Experimental Details

- Data set of morphologically complex words with different properties
- Selected from German newspaper data (13.6 million sentences)
- Morphological analysis with SMOR → select words with specific properties Schmid et al. (2004)

Netznutzungsentgelt `Netz<NN> nutzen<V>ung<SUFF><NN> Entgelt<NN>`
grid usage charge `grid<NN> use<V>age<SUFF><NN> charge<NN>`

- Prompts are formulated in German
- Evaluation: mostly semi-automatically
- Language Model: `gpt-3.5-turbo-instruct`

Noun Compounds: Predicting the Head Noun

- Compounds consisting of two nouns NN_1 NN_2 have a modifier and a head noun
- The head contains key information about the compound
  determines grammatical properties, often a more general instance of the compound: *chocolate cake → cake*
- Compare three prompts to predict the head noun
  - **P-1** What is the head noun of W?
  - **P-2** What is the head noun of the word W?
  - **P-3** What is the head noun of the compound W?
- Novel compounds: randomly generated noun-noun combinations

	prompt1	prompt2	prompt3
head found	119	270	269
modifier	96	26	21
definition	82	–	–
unsplit	1	–	4
other	2	4	6

Table 1: Results for identifying the compound head, comparing three different prompts (N=300).

	prompt1	prompt2	prompt3
head found	101	237	264
modifier	133	58	29
definition	60	1	–
unsplit	1	1	2
other	5	3	5

Table 2: Results for identifying the head of novel compounds, comparing three prompts (N=300).

- Similar outcome for existing and novel compounds
- Category "definition": attempt to explain the compound

Das Kopfnomen von Farbtreue ist die Fähigkeit, Farben korrekt und unverfälscht wiederzugeben.
The head noun of colour fidelity is the ability to correctly and accurately reproduce colours.

Derivation: Identifying Verbs in Complex Nouns

- Derivation → modification of surface forms, often different from the lemma
- Ability to derive that words are related improves generalization
- Evaluate ability to find the lemma of a verb "hidden" in a complex noun

Beschwerdebrief**beantwort**er Anfragen**beantwort**ung
answerer to complaint letters **answer**ing to requests

shared verb: beantworten (to) answer

- **P-1** What verb stem occurs in both words A and B? List the lemma.
- **P-2** What common word stem occurs in both words A and B? List the lemma.
- **P-3** What common word stem occurs in both words A and B? List the lemma and the part-of-speech.

	Prompt1	Prompt2	Prompt3
correct lemma	334	143	297
stem	7	86	8
incomplete	3	2	4
bar-word	4	1	1
noun	–	115	40
other	2	3	–

Table 4: Common verb stem task. (N=350)

Derivation: Identifying Invalid Forms

- Evaluate the model's knowledge of derivational rules
- Present the model with invalid words → identify them as not well-formed
- Set of incorrect forms: invalid combinations of productive morphemes
  Partially follow derivational rules → words are meaningless, but look like German words at a first glance:
  Constructed seemingly correctly with regard to the position and order of the added suffixes
- Test set based on *-bar adjectives* (derived from transitive verbs)

transitive vs. intransitive		
jammern + bar → jammerbar	*squeal + able → squealable*	✗
falten + bar → faltbar	*fold + able → foldable*	✓
obviously incorrect forms		
kneten + bar + keit → Knetbarkeit	*knead + able + ity → kneadability*	✓
kneten + keit + bar → knetkeitbar	*knead + ity + able → kneadityble*	✗
adding more and more morphemes		
ge + VERB + bar	*un + VERB + able*	✗
ge + VERB + bar + lich	*un + VERB + able + lich*	✗
...	...	
un+ge + VERB + bar+lich+keit	*un+ge + VERB + bar+lich+keit*	✗

Set	"yes"	"no"
*Set_intrans	19	1
Set_main	30	–
Set_main + *-keit*	30	–
*Set_main_contrastive + *-keit*	5 / 1* / 7**	15 / 2*
*Set_main + *ge-*	30	–
*Set_main + *unge-*	30	–
*Set_main + *unge-...-lich*	14	16
*Set_main + *unge-...-lichkeit*	2 / 27*	1*

Table 6: Answers to the question "*is W a word?*", ignoring the explanation part of the question. A * denotes sets with invalid words.

→ Understanding of the general meaning of *-bar* adjectivization
→ No knowledge of the underlying morphological patterns and restrictions

- *-bar*-adjectives based on intransitive verbs: incorrectly recognized as existing
- Obviously invalid words: mostly identified
- Meaningless words with stacked morphemes: mostly declared as existing
- The input word is often changed in the answer:
  *Ungeklappbarlichkeit → Ungeklappbarkeit *un-ge-collapsible-ly-ness
  Ja, das Wort "Ungeklappbarkeit" existiert und bedeutet, dass etwas nicht zusammengeklappt werden kann.
  Yes, the word "ungecollapsibleness" exists and means that something cannot be collapsed.

→ Suggests a certain insecurity

Derivation: Diminutive Forms

- Diminutive: creates a meaning of *small*: *Apfel → Äpfelchen* *apple → little apple*
- Non-concatenative operations: vowel change, addition/removal of word-final characters
- Test set consists of mostly infrequent compounds → avoid memorization effects
- Generate the word forms:

(i) given a word in diminutive form, output the form without diminutive ending
Grundschulst**ü**hlchen → Grundschulst**u**hl
elementary-school-chair

(ii) generate the diminutive form for a given word
Schönwetterw**o**lke → Schönwetterw**ö**lkchen
lovely-weather-cloud

dim → word	
correct form	235
correct (infl.)	3
head wrong	24
fuge wrong	13
incomplete	22
synonym	2
same word	1

word → dim	
correct form	213
correct (alt.)	2
head wrong	18
fuge wrong	3
incomplete	17
wrong word	7
mod_dim	17
both_dim	10
other	13

Table 7: Creating the word without diminutive form (left); creating the diminutive form (right). (N=300)

→ Correct generation for most words

- Realization of the surface form:
  wrong vowel, missing characters
- Diminutivization:
  incomplete word, morpheme in the wrong position
- Diminutive at the modifier:
  *Ausflugsort → *Ausflügchenort*
  outing-destination: nice place for outings
- Diminutive at both nouns:
  *Tannenwald → *Tännchenwäldchen* *pine-forest*

Conclusion

- Evaluation of an LLM's ability to understand morphologically complex words
- Generally good understanding of complex words
  - for example: obtain the head of a compound, generate variant of a word
  - dependent on prompt formulation
- Does the LM understand the underlying morphological patterns?
  - novel compounds: the model can handle unseen words
  - identification on invalid words: the model failed in this task → suggests a lack of understanding and knowledge
  - generation of diminutive forms: mixed results

LaTeX TikZposter