Large Language Models - Seminar

Introduction: Linguistic Concepts for NLP

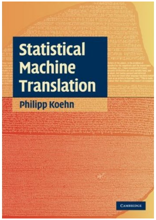
Marion Di Marco

16. April 2024

Introduction and Overview

- Linguistic concepts relevant for natural language processing
  - Words: what is a word?
  - Sentences and part of speech
  - Syntax and parse trees
  - Morphology: vocabulary and subword segmentation

The slides are partially based on Chapter 2 "Words, Sentences and Corpora" from the book "Statistical Machine Translation" (Philipp Koehn)



Outline

Words

- Word: basic atomic unit of meaning

house → 

- Adapt the meaning based on the context

  ... their parents' house ...

  ... the White House ...

- Almost all uses of *house* are connected to the basic unit of meaning
- Smaller units such as syllables or sounds (*hou* or *s*) do not evoke the mental image of *house*

What is a Word?

- Notion of words seems straightforward for English → space separated
- Some writing systems do not clearly mark words as unique units
  for example, Chinese is written without spaces between the words
- Complex words and compounding: some words appear to be one word, but consist of several parts
  - English: *homework, tumbledown, blackboard*
  - German: *Apfelkuchen (apple cake), feuerlöscherrot (fire extinguisher red)*
    Rinderkennzeichnungsfleischetikettierungsüberwachungsaufgabenübertragungsgesetz[1]
  - Finnish: *istahtaisinkohankaan (I wonder if I should sit down for a while after all)*[2]

[1] https://www.duden.de/sprachwissen/sprachratgeber/Die-langsten-Worter-im-Dudenkorpus

[2] https://en.wikipedia.org/wiki/Finnish_language

Tokenization

- For NLP tasks
  - consistent representation of the data as a sequence of tokens
  - keep the vocabulary as small as possible
- Do not blow up the vocabulary with different forms such as *house* and *house*, and *house!* and *"house"*

- Tokenization: breaking raw text into words assuming words as they appear on the surface level as tokens

- Languages with similar concepts of words than English: essentially splitting off punctuation
- Writing systems without spaces or languages with highly complex words: segmentation is more challenging

Tokenization

- Example for English tokenization
- What about
  - ... possessive markers (*Tom's*) and merged words (*doesn't*)?
  - ... abbreviations (*Abk.*) or similar items containing a dot?
  - ... hyphenation (*co-operate*)?
- Possible further normalization: lowercasing

Raw text	My son's friend, however, plays a high-risk game.
Tokenized	My son 's friend , however , plays a high @-@ risk game .
Lowercased	my son 's friend , however , plays a high @-@ risk game .

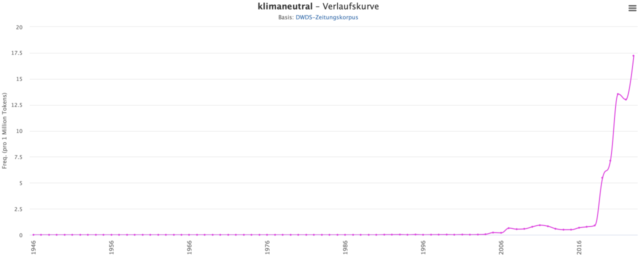
Figure 2.1 Tokenization and lowercasing: Basic data processing steps for machine translation. Besides splitting off punctuation, hyphenated and merged words may be broken up.

Figure from "Statistical Machine Translation"

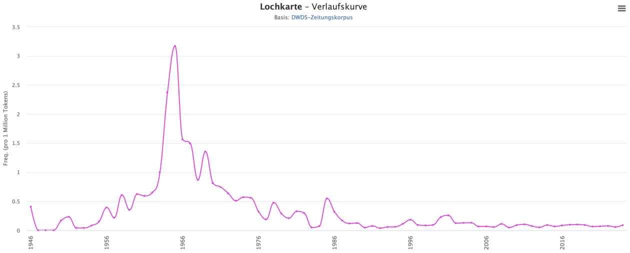
What are the Words of a Language?

New words emerge, others fall out of use:

https://www.dwds.de/r/plot



"climate neutral"



"punch card"

Corpora and Word Distribution

- The vocabulary of a language is fluid
- In practice: text corpus with a fixed set of words
- Continually update with new data → larger corpora

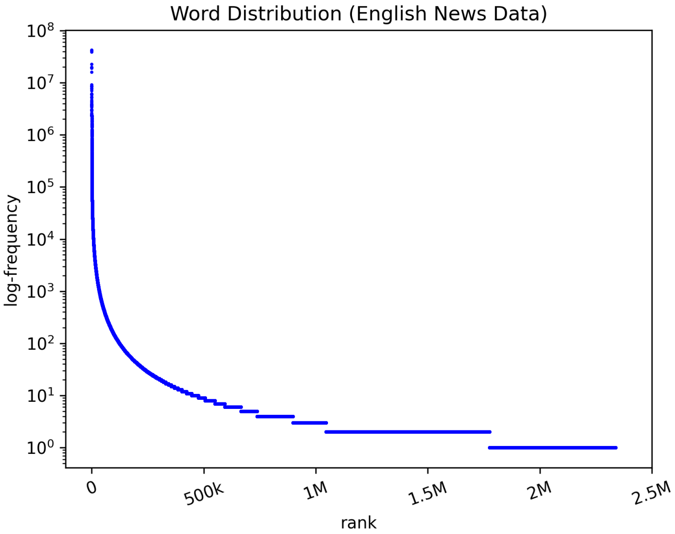
- English news data (33M sentences):

freq	**word**
42380661	,
40887715	the
38696981	.
22720213	to
19785952	and
19644063	of
19025360	a
15930678	in
9164833	's
...	...

freq	**word**
17313	timing
17304	filming
17303	overcome
17300	magic
17299	innocent
17296	admit
17278	patterns
17275	rolling
17269	formally
...	...

freq	**word**
3	yoghurt-coated
3	yesteray
3	yellow-beaked
3	worrried
3	womansplain
...	...
2	ruminococcaceae
...	...
1	north-northwestern
...	...

Corpora and Word Distribution



Outline

Parts of Speech and POS tagging

- Parts of speech: grammatical categories or word classes
- Words within the same word class: similar syntactic behaviour and similar grammatical properties

- Part-of-Speech tagging: labeling the POS tags of words in a text

- Well-established strategy:
  - annotate a large amount of text with POS-tags
  - train a tagger on the annotated data

- No trivial task:
  - words that appear the same can occur in different functions, for example *to house* (VERB) ↔ *the house* (NOUN)
  - classify previously unseen words

POS Tagging – Example

word	**POS**
When	WRB
the	DT
space	NN
shuttle	NN
was	VBD
approved	VBN
in	IN
1972	CD
,	,
NASA	NP
officials	NNS
predicted	VBD
that	IN
they	PP
would	MD
launch	VB
one	CD
every	DT
week	NN
or	CC
two	CD
.	SENT

Function Words and Content Words

Content words

- Words with lexical content
  - Nouns → refer to entities
  - Verbs → actions
  - Adjectives → attributes of entities
  - Adverbs → attributes of actions
- Open-class words

Function words

- Words with little to no lexical meaning
- Provide the structure of a sentence: express grammatical relations between content words
- For example prepositions, pronouns, articles, auxiliary verbs, ...
- Closed-class words

What does that mean for NLP applications?

- Content words:
  - continually evolving non-finite set of words
  - many existing words, with new words being introduced
  - depending on the language: further inflectional variants → morphology
- Need for large text corpora to span many topics and domains for sufficient coverage

- Function words:
  - comparatively small set of words
  - make up a large part of the overall word count
  - their interpretation is often context-dependent (for example, *that* as a determiner or relative pronoun)
  - depending in the language: different realization of linguistic concepts → morphology, sentence structure

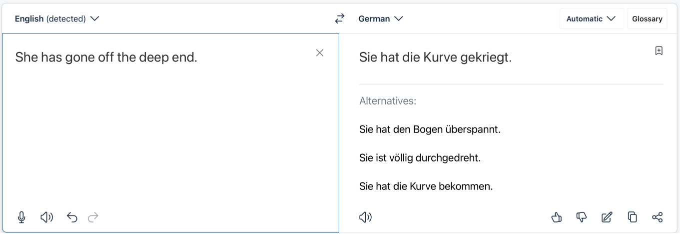
Non-compositional Phrases

- The idea of words as basic units of meaning does not always hold
- For example: idiomatic expressions
  she's gone off the deep end
  er hat nicht mehr alle Tassen im Schrank
- All words in the phrases have a distinct meaning that is not related to the meaning of the phrase (*crazy/verrückt*)
- Context: need to consider the entire phrase to derive the meaning
- Challenging for many NLP tasks
- For the sake of simplicity: assume words as the basic units of meaning

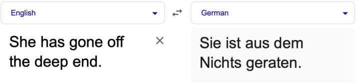
Non-compositional Phrases

- On a side note: *to go off the deep end* seems to be difficult to translate

DeepL:



Google Translate:



PONS:

Outline

Sentences

- Words → atomic units of meaning
- Sentence → combination of words following the rules of a language

(1) *Jane bought the house*

→ the **verb** *bought* is the central element

→ the verb has two arguments:
subject *Jane* and **object** *house*

(2) *Jane gave Alice a cookie.*

→ *gave/give* has three arguments:
subject *Jane* and **direct object** *cookie* and **indirect object** *Alice*

- Syntax: studies how to combine words into larger units such as phrases or sentences

Parse Trees

- Different grammar formalism to express the structure of a sentence (for example, phrase structure grammar, dependency structures, lexical functional grammar)
- Parse trees: illustrate the grammatical structure of a sentence
- Dependency structures: display relationship between words
  - one word is the head of the sentence, dependent on a notional ROOT
  - all other words are dependent on another word

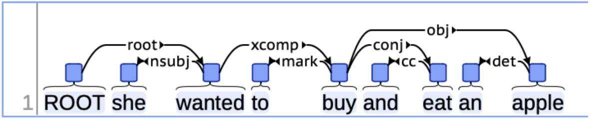


Figure from https://universaldependencies.org/u/overview/syntax.html

Syntax across Languages

- Linguistic concepts and processes are realized differently
- **Analytic languages**
  - syntactic information is mainly expressed by means of function words (e.g., prepositions, modifiers)
  - syntactic functions (subject, object) are assigned via word order
  - For example English, Norwegian, Danish
- **Synthetic languages**
  - grammatical information is synthesized into one word by means of (inflectional) morphology (e.g. grammatical case instead of prepositions)
  - relatively free word order
  - For example Slavic languages, German, Finnish, Turkish
- Often no clear distinction: languages can have features of both groups

Universal Dependency Treebank

- UDP: developing cross-linguistically consistent treebank annotation for many languages
- Tree structures for English, Bulgarian, Czech and Swedish

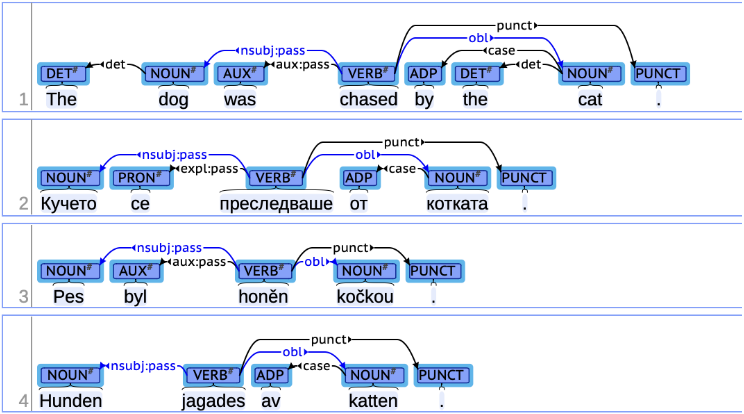


Figure from https://universaldependencies.org/introduction.html

Linguistic Structure in Large Language Models

- Language models perform very well at many language tasks
- To what extent can these abilities be attributed to generalizable linguistic understanding vs. surface-level lexical patterns?
- Can we obtain linguistic structure from LMs?

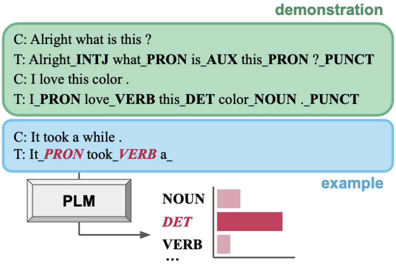


Figure 1: Sequence tagging via structured prompting. Each predicted label is appended to the context along with the next word to iteratively tag the full sentence.

Prompting Language Models for Linguistic Structure. Blevins et al. (2023)

Outline

Morphology

- Morphology: studies the internal structure and composition of words

- Inflectional morphology: addition of a morpheme, usually a suffix, to express grammatical categories
- Does not change the core lexical meaning of the words
- Some examples:
  - number: *house* → *houses*
  - tense: *machen* → *machte*

- Derivational morphology: forming a new word from existing words
- This changes the lexical interpretation of the word
- Some examples:
  - Addition of particle: *ab* + *machen* → *abmachen* (*'off make': remove*)
  - Adjectivization: $\textit{fold}_{\textit{verb}}$ + *-able* → $\textit{foldable}_{\textit{adj}}$

Morphological Complexity

- Morphologically poor languages: express relationships between words mostly with function words

- Morphologically rich languages: morphological variations
  - verbal inflection
  - nominal inflection
  - word formation processes: for example compounding
    Apfel + *Kuchen* → *Apfelkuchen* (*apple cake*)

- More morphological variation: larger vocabulary of surface forms

Example: Czech Nominal Inflection

- Inflection paradigm for the Czech adjective *mladý* (*young*)

<table>
<tr><th></th><th></th><th>Masculine animate</th><th>Masculine inanimate</th><th>Feminine</th><th>Neuter</th></tr>
<tr><th rowspan="7">Sg.</th><th>Nominative</th><td colspan="2">mladý</td><td>mladá</td><td>mladé</td></tr>
<tr><th>Genitive</th><td colspan="2">mladého</td><td>mladé</td><td>mladého</td></tr>
<tr><th>Dative</th><td colspan="2">mladému</td><td>mladé</td><td>mladému</td></tr>
<tr><th>Accusative</th><td>mladého</td><td>mladý</td><td>mladou</td><td>mladé</td></tr>
<tr><th>Vocative</th><td colspan="2">mladý!</td><td>mladá!</td><td>mladé!</td></tr>
<tr><th>Locative</th><td colspan="2">mladém</td><td>mladé</td><td>mladém</td></tr>
<tr><th>Instrumental</th><td colspan="2">mladým</td><td>mladou</td><td>mladým</td></tr>
<tr><th rowspan="7">Pl.</th><th>Nominative</th><td>mladí</td><td colspan="2">mladé</td><td>mladá</td></tr>
<tr><th>Genitive</th><td colspan="4">mladých</td></tr>
<tr><th>Dative</th><td colspan="4">mladým</td></tr>
<tr><th>Accusative</th><td colspan="3">mladé</td><td>mladá</td></tr>
<tr><th>Vocative</th><td>mladí!</td><td colspan="2">mladé!</td><td>mladá!</td></tr>
<tr><th>Locative</th><td colspan="4">mladých</td></tr>
<tr><th>Instrumental</th><td colspan="4">mladými</td></tr>
</table>

Figure from `https://en.wikipedia.org/wiki/Czech_declension`

Example: French Verbal Inflection

Inflection paradigm for the French verb *voir* (*to see*)

INDICATIF

Présent		Imparfait		Passé simple		Futur simple	
je	vois	je	voyais	je	vis	je	verrai
tu	vois	tu	voyais	tu	vis	tu	verras
il/elle/on	voit	il/elle/on	voyait	il/elle/on	vit	il/elle/on	verra
nous	voyons	nous	voyions	nous	vîmes	nous	verrons
vous	voyez	vous	voyiez	vous	vîtes	vous	verrez
ils/elles	voient	ils/elles	voyaient	ils/elles	virent	ils/elles	verront

SUBJONCTIF

Présent			Imparfait		
que	je	voie	que	je	visse
que	tu	voies	que	tu	visses
qu'	il/elle/on	voie	qu'	il/elle/on	vît
que	nous	voyions	que	nous	vissions
que	vous	voyiez	que	vous	vissiez
qu'	ils/elles	voient	qu'	ils/elles	vissent

CONDITIONNEL

Présent	
je	verrais
tu	verrais
il/elle/on	verrait
nous	verrions
vous	verriez
ils/elles	verraient

FORMES IMPERSONNELLES

Infinitif	Participe présent	Participe passé
voir	voyant	vu(e)

- In addition: composed tenses
- In contrast: *(to) see, sees, saw, seen, seeing*

Overview from `https://en.pons.com/verb-tables/french/voir`

Example: Agglutinative Languages

- Agglutination: process of forming new words by concatenating morphemes that correspond to syntactic features

Turkish	**English**
duy(-mak)	*(to) sense*
duygu	*sensation*
duygusal	*sensitive*
duygusallaş(-mak)	*(to) become sensitive*
duygusallaştırıl(-mak)	*(to) be made sensitive*
duygusallaştırılmış	*the one who has been made sensitive*
duygusallaştırılamamış	*the one who could not have been made sensitive*
duygusallaştırılamamışlardan	*from the ones who could not have been made sensitive*

Overview from Ataman et al. (2017)

Vocabulary in Large Language Models

- Large vocabulary → data sparsity
  → some forms only occur infrequently or even not at all
- Generally challenging for NLP applications

- Interpretation of a seen form:
  → what does the particular realization of a word mean?
- Generation of an appropriate form:
  → what should a form look like in the given context?

- Just add more training data?
  - more data certainly helps ...
  - ... but still puts morphologically rich languages at a disadvantage
- Ideally: generalization

Vocabulary in Large Language Models

- Language models are trained on huge amounts of data, often on multilingual training data
- For practical reasons: vocabulary needs to be capped
- Pre-trained language models typically rely on sub-word units
  - handle unknown words
  - for better efficiency due to reduced
- Example from ChatGPT:

```
Many words map to one token, but some don't: indivisible.

The Nile crocodile (Crocodylus niloticus) is a large crocodilian native
 to freshwater habitats in Africa. It is widely distributed in sub
-Saharan Africa.

Das Nilkrokodil ist das größte Krokodil Afrikas und erreicht normaler
weise Längen von 3 bis 4 m.
```

Vocabulary and Sub-word Units

- Subword units are often based on WordPiece or BPE

  Sennrich et al. (2016)

- Frequency-based compression algorithms:
  - start with small vocabulary (character-level)
  - iteratively merge the most common tuples
    until desired vocabulary size is reached
  - keep frequent words intact, segment less frequent ones

- Example: playing → `play ##ing`

- Is this always a good idea?
- What about languages with more complex morphology?

Vocabulary and Sub-word Units

- Segmentation based on BPE or WordPiece is not linguistically guided
- Resulting sub-words are not always meaningful linguistic units
- `mitternacht|s|blau(e|en|s)`

  the/a midnight blue car(s)

```
das mitternachtsblaue Auto.
die mitternachtsblauen Autos.
ein mitternachtsblaues Auto.
```

- Generalization issues:
  - the inflected word part *blau* (*blue*) is represented differently
  - the split does not adhere to morpheme boundaries/inflectional suffix
- Non-concatenative morphological processes cannot be captured
  - for example Umlautung: $Apfel_{Sg} \rightarrow Äpfel_{Pl}$ (*apple(s)*)

Vocabulary and Sub-word Units

- English is an analytic language without rich morphology; segmentation with WordPiece or BPE functions reasonably well
- Frequency-based segmentation is not optimal for morphologically rich languages (e.g. Arabic, Hebrew, Finnish, Turkish, ...)

Klein and Tsarfaty (2020)

- Studies for several languages: linguistically-guided segmentation in combination with frequency-based segmentation is better
  - Language modeling, machine translation

Outline

Linguistic Information in Large Language Models

- Large Language Models: state-of-the-art performance on many tasks
- Typically trained without explicit linguistic information, just large quantities of (multilingual) text

- How do LMs understand language?
  - Linguistic structure: syntax, morphology
  - Cross-lingual competence of LMs
  - World knowledge
  - Reasoning and problem solving

- Languages:
  - Most LLMs are English-centered
  - What about low-resourced languages?

Seminar Outline

- Lectures in the first part of the seminar: technical background of LMs
- Paper presentations in the second part of the seminar: discussing different topics in current papers
- For next week: please read
  Jurafsky and Martin, chapter 3: n-grams
  `https://alexfraser.github.io`

References

- Rico Sennrich, Barry Haddow, Alexandra Birch (2016):
  Neural Machine Translation of Rare Words with Subword Units
  In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics.
- Stav Klein, Reut Tsarfaty (2020):
  Getting the ##life out of living: How Adequate Are Word-Pieces for Modelling Complex Morphology?
  In Proceedings of SIGMORPHON.
- Terra Blevins, Hila Gonen, Luke Zettlemoyer (2023):
  Prompting Language Models for Linguistic Structure. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics.