n-Gram Models of Morphological Derivations

Aniello De Santo and Alëna Aksënova

Department of Linguistics @ SBU

Contact Information:
Department of Linguistics
Stony Brook University
Stony Brook, NY 11794-4376

Email:
alena.aksenova@stonybrook.edu
aniello.desanto@stonybrook.edu

Abstract

In this poster, we show two different challenges to the "standard" n-gram approach to the representation of the structure of words, and show a possible solution that allows to capture ambiguity as well as shorten the description length. Under the derivational representation of the string, one can see scope relations, as well as allow or prohibit certain combinations of affixes.

Introduction

Computational linguistics is concerned with modeling natural language, and modeling the structure of words – *morphology* – is an essential part of it. Morphology studies what different morphemes mean, how they apply to the stem, and so on. For example, *-ed* is a morpheme that can apply to a root node *walk* in order to derive its past tense *walked*.

Here we present a modification of the traditional n-gram approach to modeling language that allows the following:

- capture ambiguity, in cases such as *unlockable*;
- more efficiently limit behavior of morphemes.

This will result in better semantic extraction while parsing, and will improve our understanding of possible morpheme combinations.

1 n-Grams Models in Natural Language

Widely known in natural language processing, n-gram models can be used to decide whether a word belongs to a language or not, by banning specific substrings of length n listed in a grammar.

Example 1: Intervocalic s voicing in German

- in GERMAN, /s/ is realized as [z] in-between two vowels:
  - (1) Faser → fa[z]er 'fiber'
  - (2) reisen → rei[z]en 'to.travel'
- other consonants are unaffected:
  - (3) Wasser → wa[ss]er 'water'
  - (4) reiste → rei[s]te 'traveled'
- banned 3-grams: {*ase, *ise, *ese, *isi, ...}

ok r e i z e n   * r e $\boxed{\text{i s e}}$ n

Example 2: Word-final devoicing in German

- in GERMAN, /d/ is realized as [t] at the end of the word:
  - (5) Kind → kin[t] 'child'
  - (6) Kinder → kin[d]er 'children'
- banned 2-grams: {*d⋉ }
- ⋊, ⋉ mark the left and right edge of a word

ok ⋊ k i n t ⋉   * ⋊ k i n $\boxed{\text{d ⋉}}$

2 Limits

A surprising number of natural language patterns can be captured by these very simple models. But **not all patterns are this well-behaved**.

A first outlier: Semantic ambiguity

- The combination of distinct morphemes can lead to multiple meanings, often expressed by the same *form*:

un-lock-able

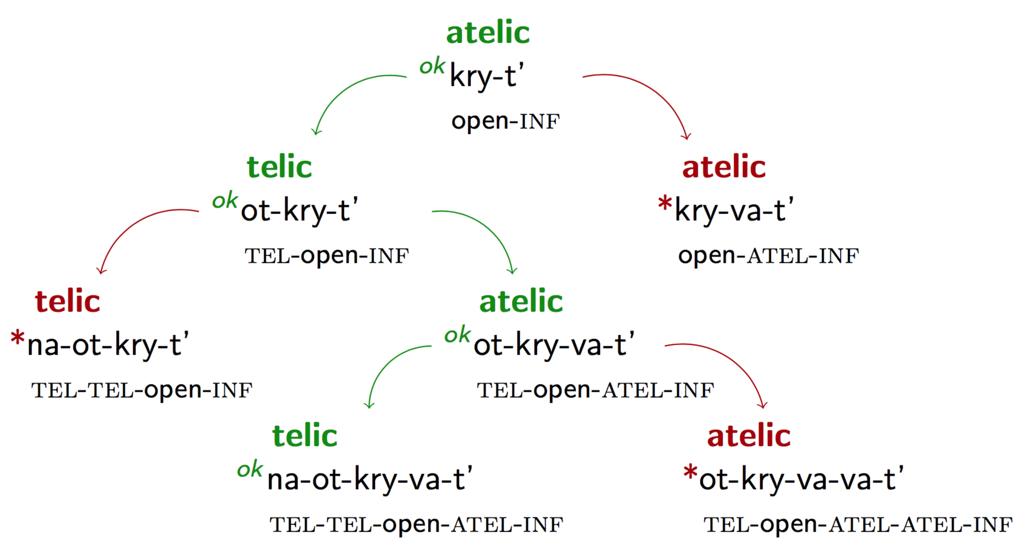
[un-lock]-able — # possible to unlock
un-[lock-able] — # not possible to lock

- n-gram models evaluate the form of the string;
- how can they account for this phenomenon?

A second outlier: Russian nominalization

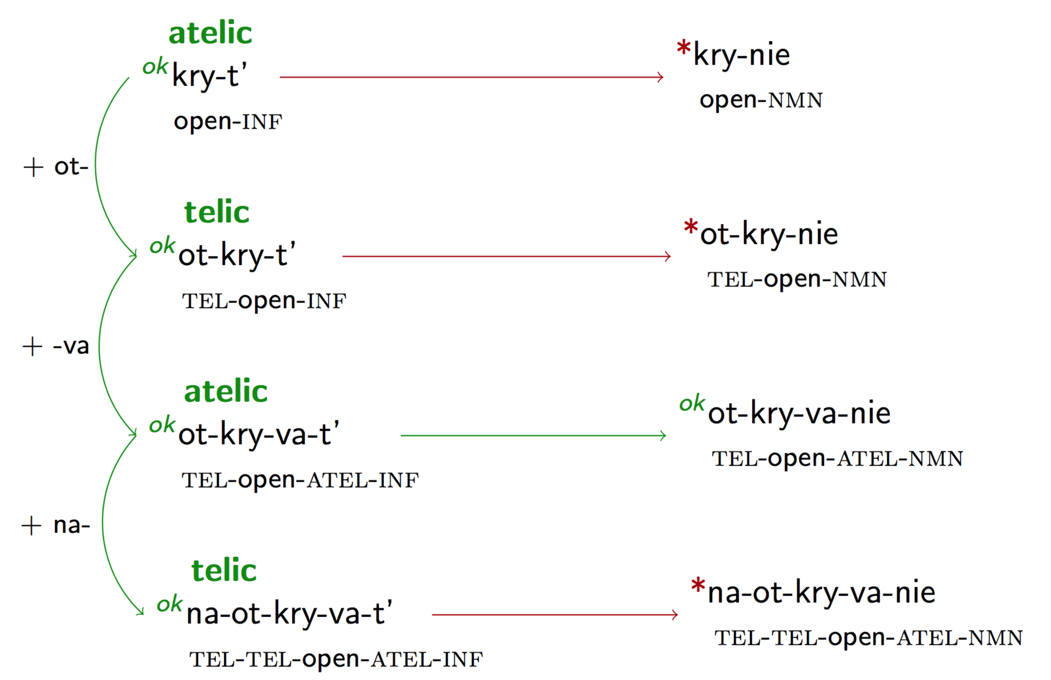
Telic-Atelic alternation:

- stems are intrinsically atelic;
- telic prefixes and atelic suffix;
- telic prefix can be added only to the atelic form;
- atelic suffix can be added only to the telic form.



The nominalization suffix *-nie*:

- cannot apply directly to the stem;
- cannot apply to a telic form;
- can only be applied after the stem is converted to an atelic form.



Problems:

- they can't capture the semantic ambiguity caused by different order of affix application;
- to capture the Russian pattern we need a combination of 5-grams (at least) → for big values of n (i.e. ≥ 3) an n-gram based analysis of the pattern gets significantly complex.

3 Solution: Derivational Strings

Idea: instead of evaluating the overt form, evaluate the ordered sequence of performed operations (*derivation*).

Back to semantic ambiguity

un-lock-able

[un-lock]-able — # possible to unlock — `stem-un-able`
un-[lock-able] — # not possible to lock — `stem-able-un`

The derivational representation captures the semantic ambiguity caused by different order of affix application.

Being able to access the history of how a word was formed significantly simplifies the problem of semantic ambiguity. Can we use this realization to analyze the Russian nominalization pattern?

Back to Russian nominalization

First of all, we need to encode the derivational history of nominalized words:

open-INF `stem-inf` → *open-NMN `stem-nmn`

TEL-open-INF `stem-tel-inf` → *TEL-open-NMN `stem-tel-nmn`

TEL-open-ATEL-INF `stem-tel-atel-inf` → TEL-open-ATEL-NMN `stem-tel-atel-nmn`

TEL-TEL-open-ATEL-INF `stem-tel-atel-tel-inf` → *TEL-TEL-open-ATEL-NMN `stem-tel-atel-tel-nmn`

We can now give a very simple n-gram account. What kind of substrings should our model ban?

- *`tel-nmn`: the telic form cannot be nominalized;

  *`na-ot-kry-va-nie`
  `stem-tel-atel-tel-nmn`

- *`stem-nmn`: prohibit nominalization of the verbal root.

  *`kry-nie`
  `stem-nmn`

Russian nominalization: derivational n-grams

- Banned n-grams: { *tel-nmn, *stem-nmn }

ok ot-kry-va-nie	*na-ot-kry-va-nie
TEL-open-ATEL-NMN	TEL-open-ATEL-TEL-NMN
`stem-tel-atel-nmn`	`stem-tel-atel-tel-nmn`
$\boxed{\text{ot - kry - va - nie}}$	stem - tel - $\boxed{\text{atel - nmn}}$
$\boxed{\text{na - ot - kry - va - nie}}$	stem - tel - atel - $\boxed{\text{tel - nmn}}$
Overt strings: 5-grams	**Derivational strings: 2-grams**

Conclusion

Here we presented two cases that are potentially problematic for n-gram approach to morphology. The first issue is posed by ambiguous words such as *unlockable*, where different orders of morpheme application result in different interpretations. Another problem is Russian nominalization, which includes seemingly non-local processes.

To solve these problems, we proposed to apply n-grams to *derivational* strings encoding the order of operations performed in forming a word. Such representation of a word:

- allows to extract scope relations caused by different order of morpheme attachments;
- allows for local descriptions of dependencies that are non-local in the overt form.

This approach improves meaning extraction and makes sound predictions about possible combinations of morphemes, while the model still remains cognitively plausible and highlights linguistic generalizations.

References

- **Aksënova**, A., T. Graf, and S. Moradi. 2016. Morphotactics as Tier-based Strictly Local Dependencies. In *Proceedings of SIGMORPHON 14*, 121–130.
- **Aksënova**, A and A. De Santo. 2017. Strict Locality in Morphological Derivations. In *Proceedings of CLS 53*.
- **Beesley**, K., and L. Kartunnen. 2003. *Finite state morphology*. Stanford, CA: CSLI Publications.
- **Graf**, T., and J. Heinz. 2016. *Tier-based strict locality in phonology and syntax*. Ms., Stony Brook University and University of Delaware.
- **Heinz**, J. and Idsardi W. 2013. What Complexity Differences Reveal About Domains in Language. *Topics in CogSci* 5(1):111–131.
- **Kaplan**, R. M., and M. Kay. 1994. Regular models of phonological rule systems. *CL* 20:331–378.
- **Pazelskaya**, A. 2012. Verbal prefixes and suffixes in nominalization: grammatical restrictions and corpus data. In *The Russian Verb*, 245–261.

Acknowledgements

We are grateful to Thomas Graf, Jeff Heinz, and the participants to the 53rd Meeting of the Chicago Linguistics Society for very useful comments and suggestions.