UNLEMMATIZATION

RECOVERING WORD FORMS IN MORPHOLOGICALLY RICH LANGUAGES

Anton Alekseev, Sergey Nikolenko

September 22, 2017

Steklov Institute of Mathematics at St. Petersburg

INTRODUCTION

Given a lemmatized sentence, it is possible for a human to suggest word forms so that the sentence starts making sense.

дарио оставлять жена в панама и предпринимать путешествие в столица аргентина

- *дарио оставил жену в панаме и предпринял путешествие в столицу аргентины*
- *дарио оставляет жён в панаме и предпринимает путешествие в столицу аргентины*

This “decoding task” of suggesting a set of word forms so that the sentence would become grammatical (and ultimately meaningful) could be modeled using recent developments in machine learning.

INTRODUCTION

The two natural variations are

(1) finding all possible “unlemmatizations”, probably in the form of a list of most probable paths through a word network, similar to speech recognition;

(2) generating only one possible answer that humans would not complain about.

INTRODUCTION

The two natural variations are

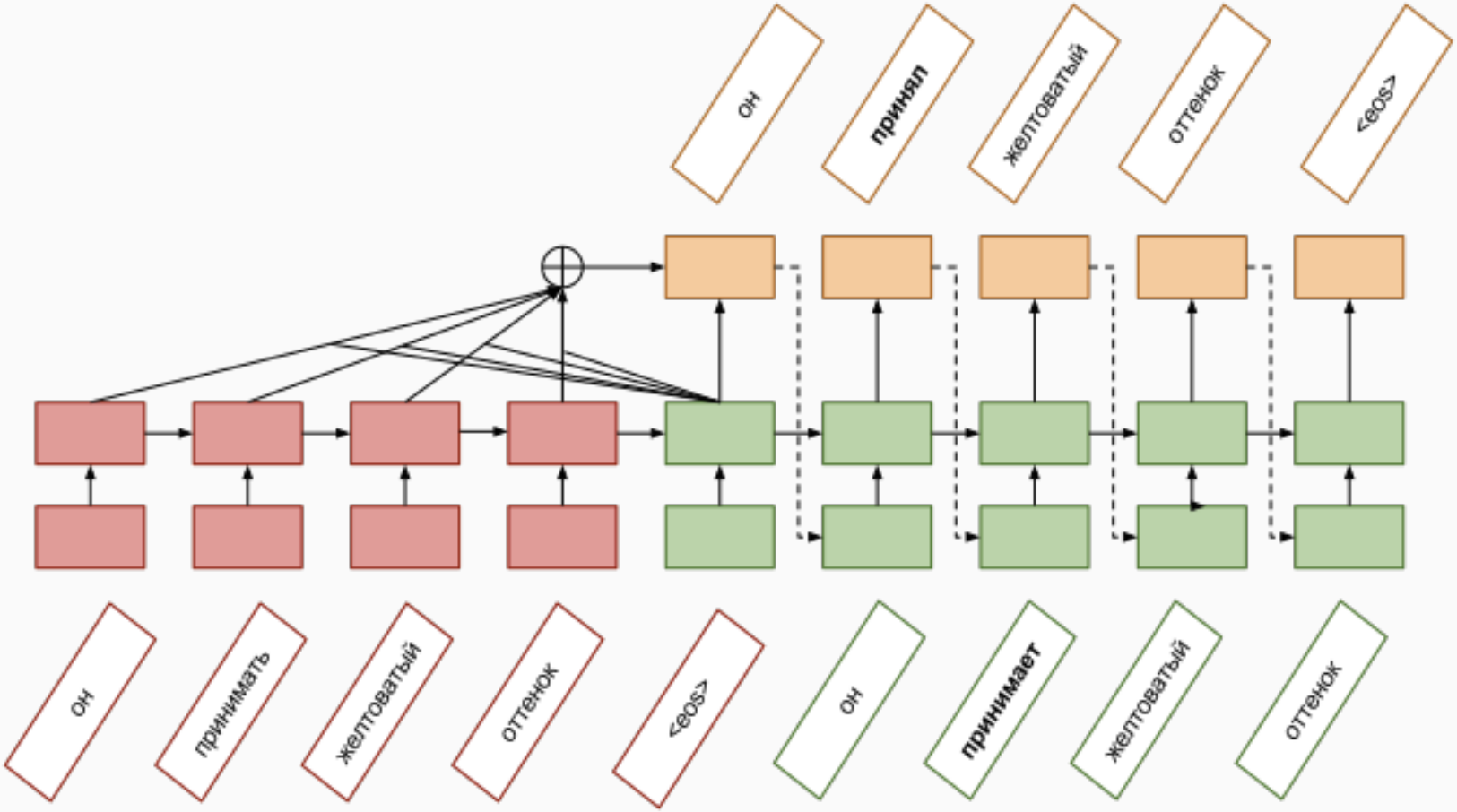
(1) ~~finding all possible "unlemmatizations", probably in the form of a list of most probable paths through a word network, similar to speech recognition;~~

(2) generating only one possible answer that humans would not complain about.

POSSIBLE APPLICATIONS

- In native advertising
- NLP systems output postprocessing
- Data augmentation in statistical NLP
- More ideas?

FIRST APPROACH: OPEN-NMT + RU-WIKIPEDIA



CHERRY-PICKED RESULTS

	text
lemmatized	в тот же год занимать пост министр внутренний дело
true answer	в том же году занимает пост министра внутренних дел
wiki-10k	в том же году занял принял франции письмо государства
wiki-100k	в том же году занял пост министра внутренних дел
lemmatized	после обработка он принимать зеленоватый оттенок
true answer	после обработки он принимает зеленоватый оттенок
wiki-10k	после завершения он принял хозяйство львовуголь
wiki-100k	после обработки он принял зеленоватый оттенок

CHALLENGES

BLEU expects matching word by word, however, different word forms sometimes should not be treated as an error

Possible solutions — proxies from other NLP models

“grammaticality” could be estimated using the probabilities of generated sentences computed

(1) by some gold standard language model,
(2) by a probabilistic POS-tagging model.
(3) more ideas?

FUTURE WORK DIRECTIONS

- Apply statistical machine translation methods
- Do a thorough analysis of errors and find their possible causes
- Apply character-level machine translation models
- Apply SOTA MT models, e.g. *Attention is All You Need*
- Construct a **large training corpus** based on Russian fiction texts or user-generated texts from social networks
- Sequence learning methods with grammatical information as a hidden state

Thank you for your attention!

Questions?

Welcome to our poster stand