

COMP90042

Web search and text analysis

Workshop Week 7

xudong.han@unimelb.edu.au

https://github.com/HanXudong/COMP90042_Workshops

entity abstraction... communication message...	entity abstraction... psychological... cognition...	entity abstraction... communication message... statement pleading charge... accusation...	entity abstraction... group... collection...	entity abstraction... measure system of meas... information meas...
information				
	entity physical... process... processing data process... operation computer op...	entity abstraction... psychological... cognition... process... basic cog... memory...	entity abstraction... psychological... event act...	
retrieval				

information is more similar to the word *retrieval* or the word *science*

$$WuP_sim(c_1, c_2) = \frac{2 \times depth(LCS(c_1, c_2))}{depth(c_1) + depth(c_2)}$$

entity
abstraction
psychological
cognition
content
knowledge domain
discipline

entity
abstraction
psychological
cognition
ability

Science

entity
abstraction...
communication
message...

entity
abstraction...
psychological...
cognition...

entity
abstraction...
communication
message...
statement
pleading
charge...
accusation...

entity
abstraction...
group...
collection...

entity
abstraction...
measure
system of meas...
information meas...

information

information

1

2

3

4

5

1

0.30

0.61

0.23

0.30

0.28

science

2

0.36

0.72

0.27

0.36

0.33

At the end of this tutorial you will be able to...

1. explain the main ideas of several common POS tagging approaches
2. explain key differences and similarities between N-gram language model and feed-forward neural language models.
3. explain the basic meaning of RNN and its advantages over the feed-forward model.

Q1 What is a POS tag

- A part of speech (abbreviated form: PoS or POS) is a category of words which have similar grammatical properties.

Tag	Description	Example	Tag	Description	Example	Tag	Description	Example
CC	coordinating conjunction	<i>and, but, or</i>	PDT	predeterminer	<i>all, both</i>	VBP	verb non-3sg present	<i>eat</i>
CD	cardinal number	<i>one, two</i>	POS	possessive ending	<i>'s</i>	VBZ	verb 3sg pres	<i>eats</i>
DT	determiner	<i>a, the</i>	PRP	personal pronoun	<i>I, you, he</i>	WDT	wh-determ.	<i>which, that</i>
EX	existential 'there'	<i>there</i>	PRP\$	possess. pronoun	<i>your, one's</i>	WP	wh-pronoun	<i>what, who</i>
FW	foreign word	<i>mea culpa</i>	RB	adverb	<i>quickly</i>	WP\$	wh-possess.	<i>whose</i>
IN	preposition/ subordin-conj	<i>of, in, by</i>	RBR	comparative adverb	<i>faster</i>	WRB	wh-adverb	<i>how, where</i>
JJ	adjective	<i>yellow</i>	RBS	superlatv. adverb	<i>fastest</i>	\$	dollar sign	<i>\$</i>
JJR	comparative adj	<i>bigger</i>	RP	particle	<i>up, off</i>	#	pound sign	<i>#</i>
JJS	superlative adj	<i>wildest</i>	SYM	symbol	<i>+, %, &</i>	“	left quote	<i>‘ or “</i>
LS	list item marker	<i>1, 2, One</i>	TO	“to”	<i>to</i>	”	right quote	<i>’ or ”</i>
MD	modal	<i>can, should</i>	UH	interjection	<i>ah, oops</i>	(left paren	<i>[, (, {, <</i>
NN	sing or mass noun	<i>llama</i>	VB	verb base form	<i>eat</i>)	right paren	<i>],), }, ></i>
NNS	noun, plural	<i>llamas</i>	VBD	verb past tense	<i>ate</i>	,	comma	<i>,</i>
NNP	proper noun, sing.	<i>IBM</i>	VBG	verb gerund	<i>eating</i>	.	sent-end punc	<i>. ! ?</i>
NNPS	proper noun, plu.	<i>Carolinas</i>	VBN	verb past part.	<i>eaten</i>	:	sent-mid punc	<i>: ; ... - -</i>

Tagged text Example

The/DT limits/NNS to/TO legal/JJ absurdity/NN
stretched/VBD another/DT notch/NN this/DT week/NN
when/WRB the/DT Supreme/NNP Court/NNP
refused/VBD to/TO hear/VB an/DT appeal/VB from/IN
a/DT case/NN that/WDT says/VBZ corporate/JJ
defendants/NNS must/MD pay/VB damages/NNS
even/RB after/IN proving/VBG that/IN they/PRP
could/MD not/RB possibly/RB have/VB
caused/VBN the/DT harm/NN ./.

Q1a

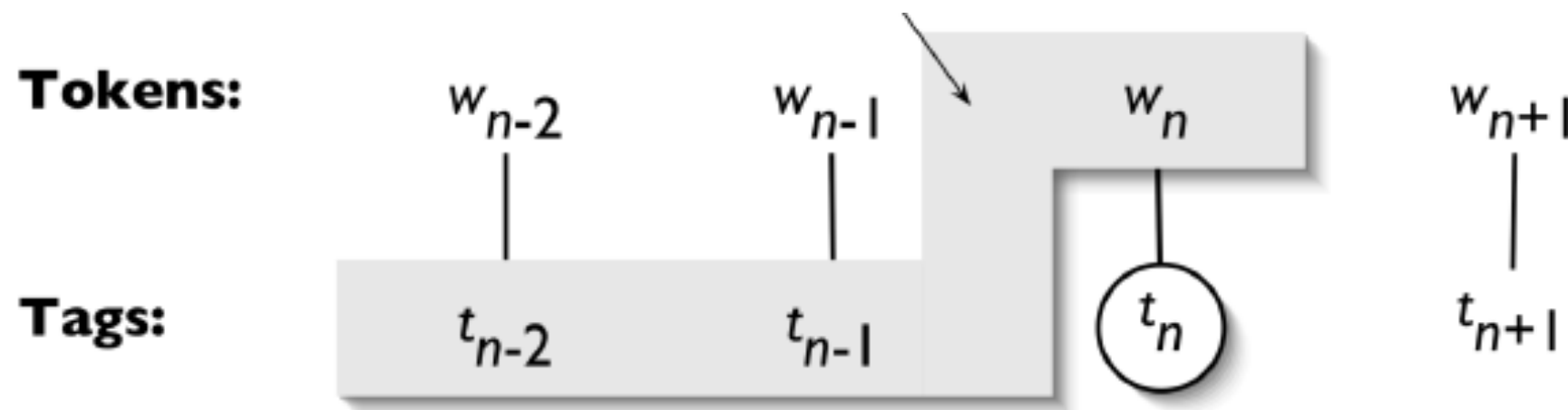
What are some common approaches to POS tagging? What aspects of the data might allow us to predict POS tags systematically?

- N-gram

- Rule-based

<https://learnenglish.britishcouncil.org/english-grammar>

- Classifier



- HMM

Q1b

Pierre Vinken, 61 years old, will join the board as a nonexecutive director Nov. 29.

Pierre	Vinken	,	61	years	old	,	will	join	the	board	as	a	nonexecutive	director	Nov	29	.
		,				,											.

- NN sing or mass noun *llama*
- NNS noun, plural *llamas*
- NNP proper noun, sing. IBM
- VB verb base form *eat*
- JJ adjective *yellow*
- MD modal *can, should*
- CD cardinal number *one, two*
- DT determiner *a, the*
- IN preposition/ subordin-conj *of, in, by*

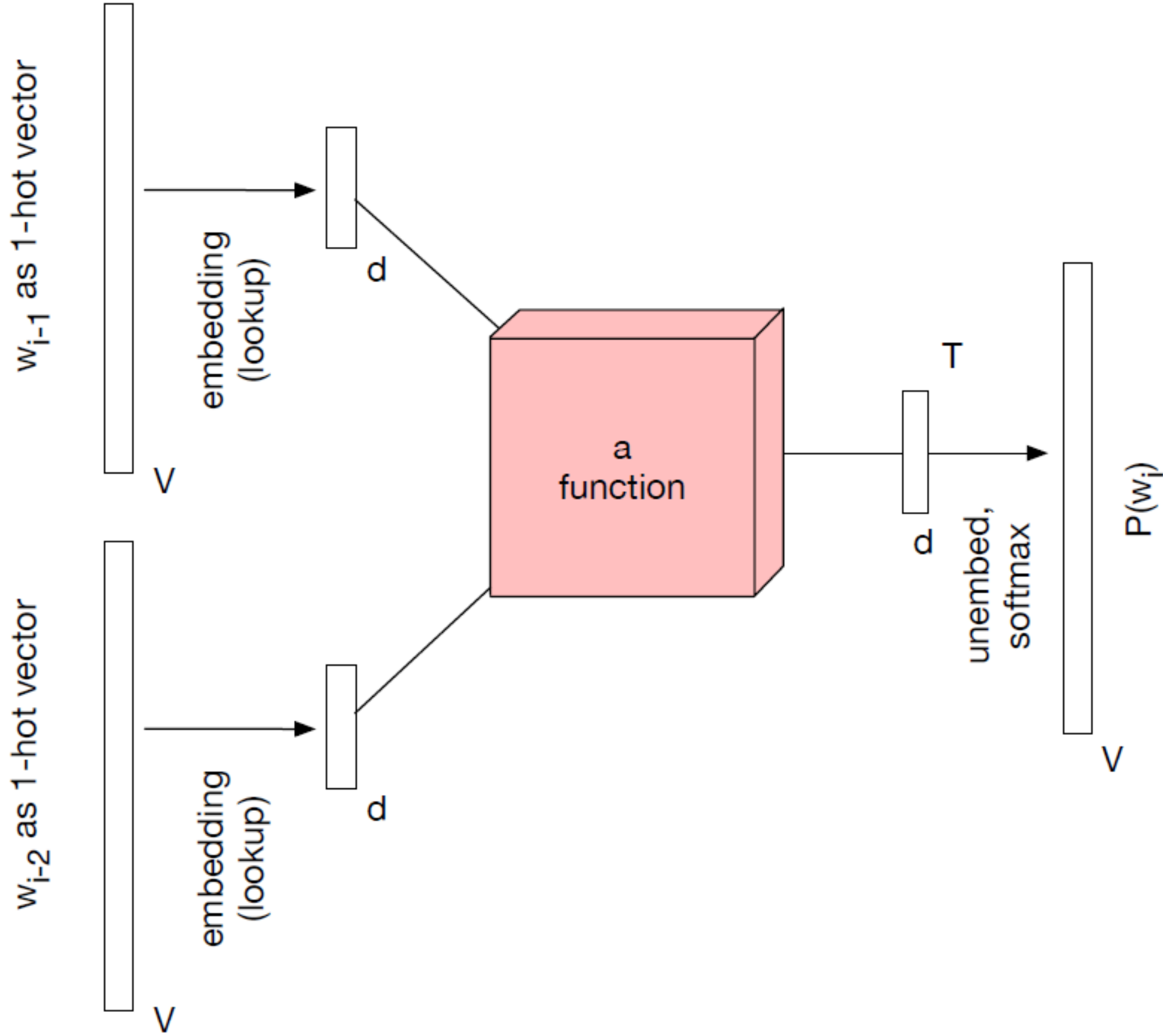
Q2

Name the key differences and similarities between n-gram language models versus feed-forward neural language models.

$$P_{add1}(w_i | w_{i-2} w_{i-1}) = \frac{C(w_{i-2} w_{i-1} w_i) + 1}{C(w_{i-2} w_{i-1}) + V}$$

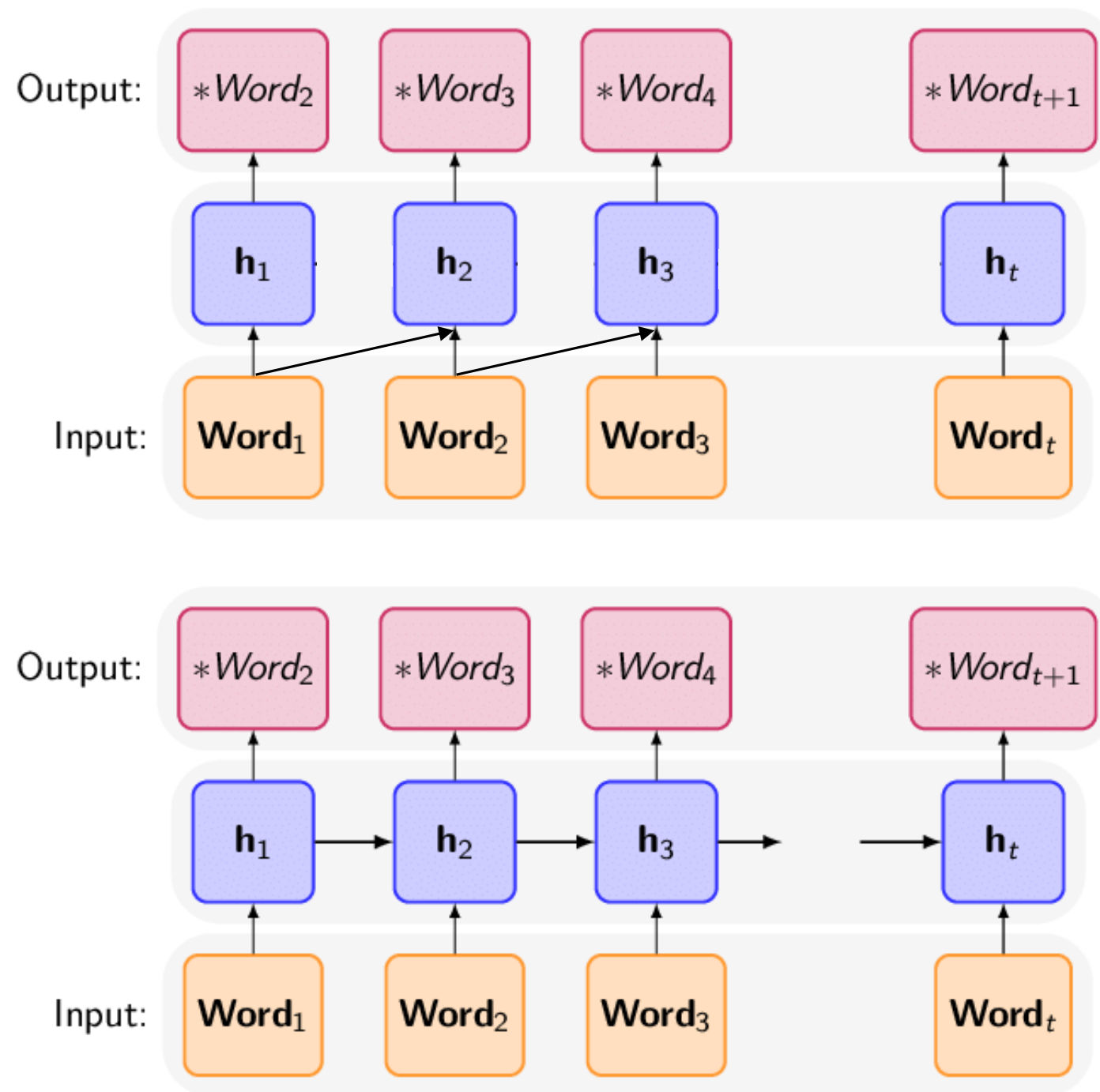
$$P(w_1, w_2, \dots, w_m) = \prod_{i=1}^m P(w_i | w_{i-2} w_{i-1})$$

Feed forward neural net LM



Q3

What does recurrent mean in the context of a recurrent neural network (RNN) language model? How does the approach differ from a feed-forward language model?



Q4

What advantage does a RNN language model have over a feed-forward language model?

- RNNLM can capture long-distance dependencies, while FFLM cannot. For example, it can balance quotes and brackets over long distances.
- (..... (..... (..)))