



HAL
open science

Bridging Worlds: The Splicing of MDD and GPT for Constrained Text Generation

Alexandre Bonlarron, Aurelie Calabrese, Pierre Kornprobst, Jean-Charles
Régis

► **To cite this version:**

Alexandre Bonlarron, Aurelie Calabrese, Pierre Kornprobst, Jean-Charles Régis. Bridging Worlds: The Splicing of MDD and GPT for Constrained Text Generation. CNIA 2023 - Conférence Nationale en Intelligence Artificielle, Association Française pour l'Intelligence Artificielle, AFIA, Jul 2023, Strasbourg, France. hal-04217503

HAL Id: hal-04217503

<https://inria.hal.science/hal-04217503>

Submitted on 26 Sep 2023

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Copyright

Bridging Worlds: The Splicing of MDD and GPT for Constrained Text Generation

PRÉSENTÉ PAR ALEXANDRE BONLARRON^(1,2)

(1) Université Côte d'Azur, Inria, France

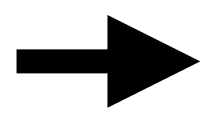
(2) Université Côte d'Azur, I3S, France

Constrained text : Standardized...

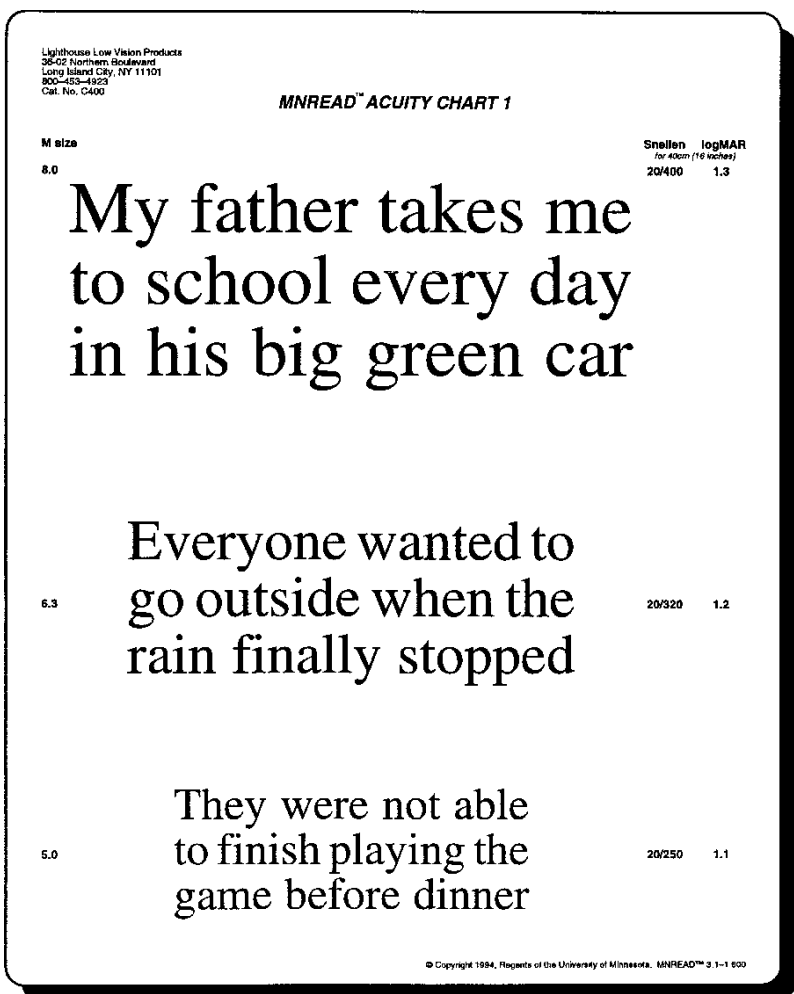
- **Standardized text** : sentences read at the same speed
- **Goal**: assess reading performance
- **How to** : obey a set of strict rules

Nous pouvons faire
une centaine de pas
dans cette direction

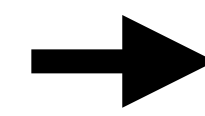
A standardized sentence (FR)



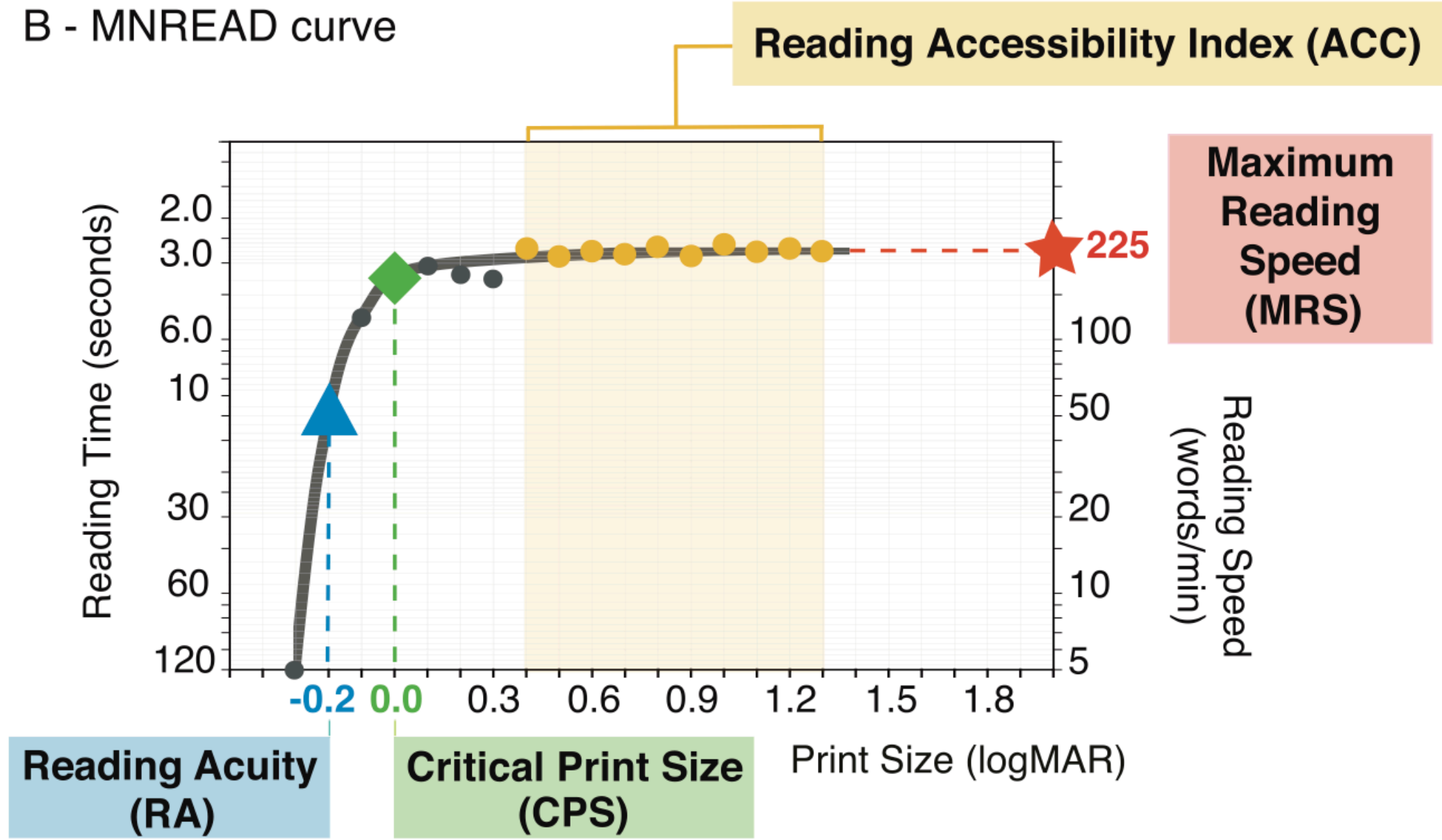
(Test MNREAD)



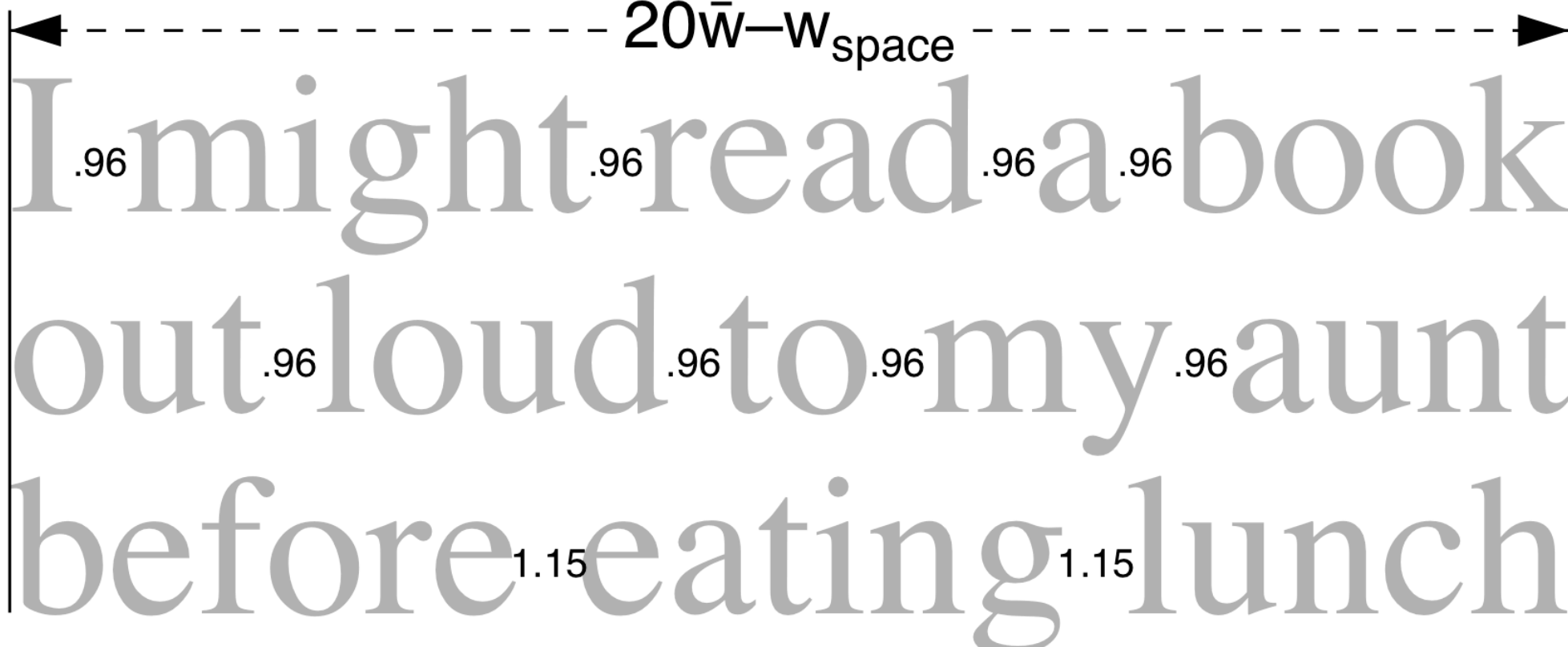
(Mansfield et al., 1993)



B - MNREAD curve

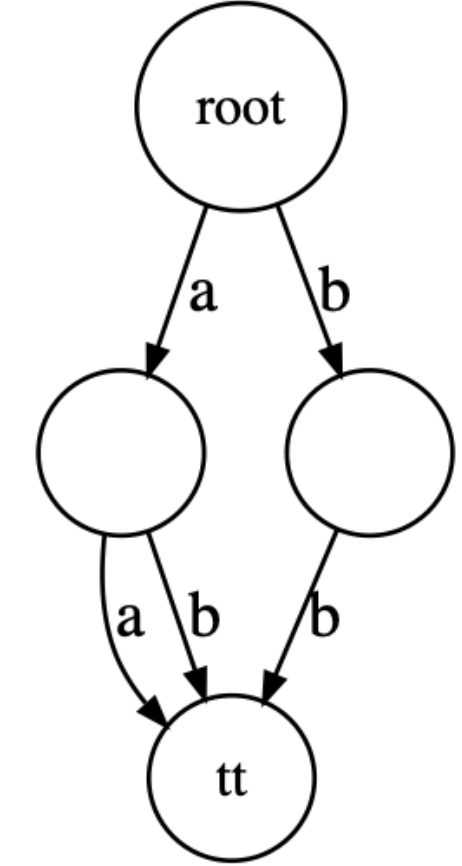


Constrained text generation



Strict text rules

+



=

Multi-Valued Decision Diagram

(MDD)

Test MNREAD

M size	Text	Score	logMAR
4.0	My father asked me to help the two men carry the box inside	20200	1.0
3.2	Three of my friends had never been to a circus before today	20100	0.9
2.5	My grandfather has a large garden with fruit and vegetables	20125	0.8
2.0	He told a long story about ducks before his son went to bed	20100	0.7
1.6	My mother loves to hear the young girls sing in the morning	20800	0.6
1.3	The young boys held their hands high in each question at school	20400	0.5
1.0	My brother asked a lot of questions but after lunch	20500	0.4
0.8	Life was wonderful and he was very happy	20400	0.3
0.6	My mother was very kind and she was very kind	20300	0.2
0.5	My mother was very kind and she was very kind	20200	0.1
0.4	My mother was very kind and she was very kind	20200	0.0
0.3	My mother was very kind and she was very kind	20200	-0.1
0.2	My mother was very kind and she was very kind	20200	-0.2
0.1	My mother was very kind and she was very kind	20200	-0.3
0.0	My mother was very kind and she was very kind	20200	-0.4

(Mansfield et al., 1993)

MDDs have already been successfully used to generate **music** and **poetry**.

It's a problem dominated by rules (**constraints**)

Focus on MNREAD phrases

PLAN

PROBLEM ★

STATE OF THE ART

OUR APPROACH

RESULT

CONCLUSION

MNREAD Rules

Grammatical rules

E.g. no punctuation

Length rules

E.g. 60 characters, between 9 and 15 words.

Display rules

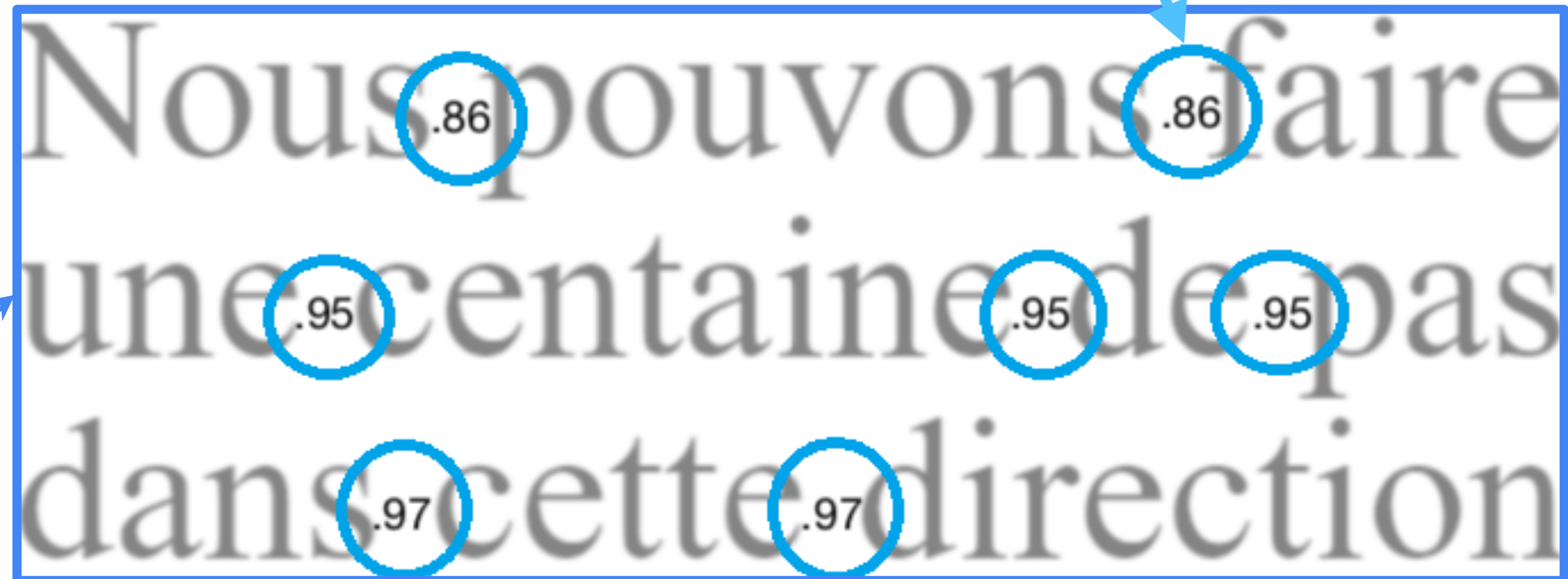
E.g. Entering the rectangle

Lexical rules

3000 words from CE2 textbooks

Example of an MNREAD sentence.

SPACE SIZE



i *There are 38 MNREAD phrases in French*

Questions :

① There are 38 MNREAD phrases in French.

Are there enough sentences?

No, a few thousand sentences are needed to detect and monitor visual pathology throughout life.

Is it really difficult to have more sentences that respect the rules ?

Naive method

Search for MNREAD type sentences in books.



Problem : This method does not scale up

Solution : We have to generate them, but how ?

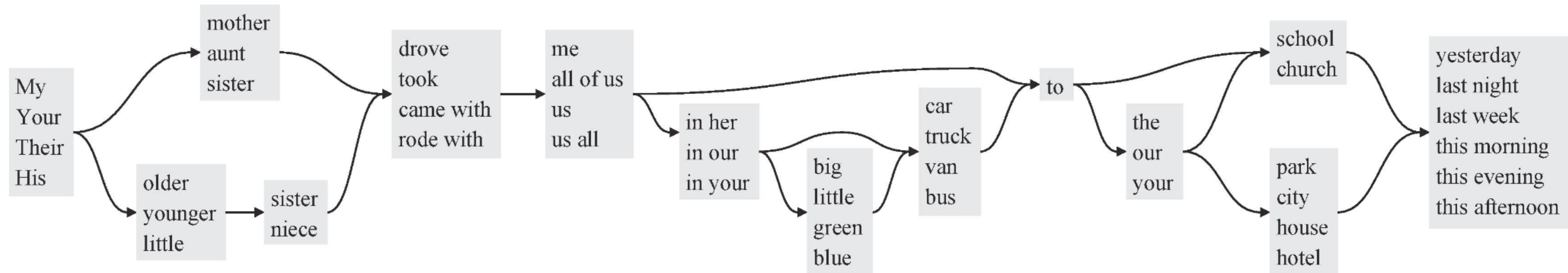
- Grammar
- Length
- Display
- Lexicon

How to generate standardized sentences?

3 method classes

LLM-based approach (GPT, BERT) + SEARCH := good text quality, but unlikely to find in an instance that satisfies the constraints.

Ad hoc method: a recent method proposed by the creators of MNREAD and based on hand-defined models (Mansfield et al., 2019).



- One "good" sentence out of 8000 (only)
- A semi-automatic method for the English language
- Non-trivial extension to Latin languages! (e.g. French?)

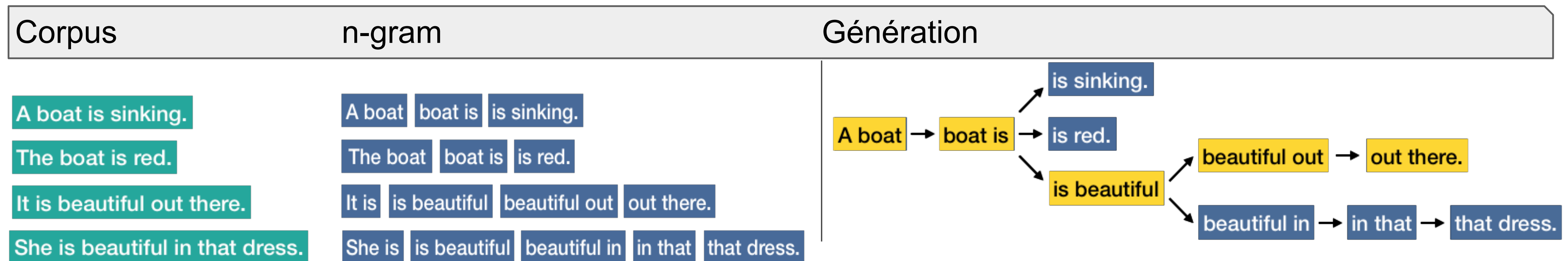
How to generate standardized sentences?

3 method classes

LLM-based approach (GPT, BERT) good text quality, but unlikely to result in an instance that satisfies the constraints.

Ad hoc method: a recent method proposed by the creators of MNREAD and based on hand-defined models (Mansfield et al., 2019).

Combinatorial optimization (n-grams based) (Papadopoulos et al., 2015)



How to integrate constraint in this scheme ?

PLAN

PROBLEM

STATE OF THE ART

OUR APPROACH ★

RESULT

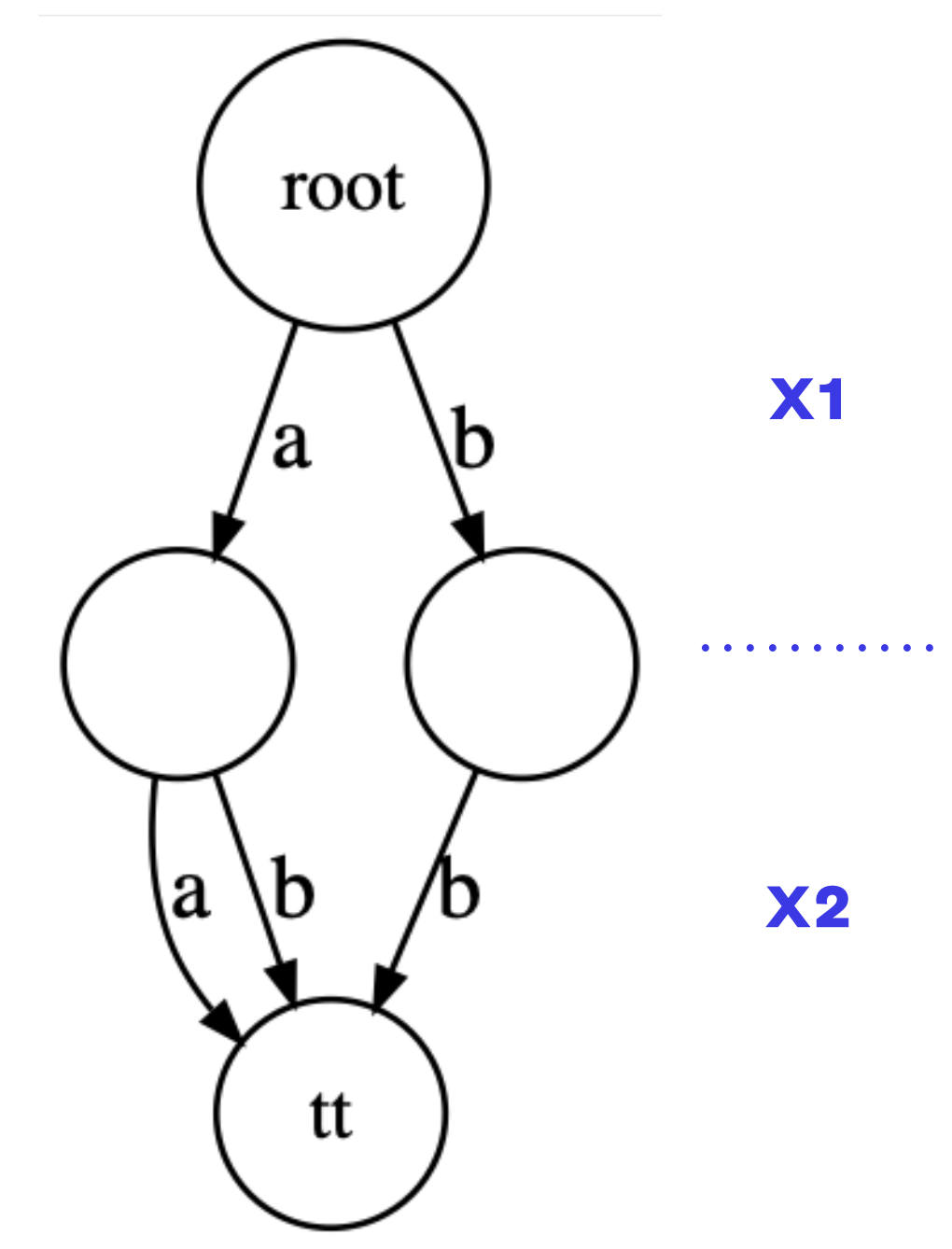
CONCLUSION

Multi-Valued Decision Diagram (MDD)

- A generalization of Binary Decision Diagrams (BDD)
- Each layer represents a variable
- Each path between root and tt is a valid assignment of the variables
- An MDD models all tuples that satisfy a constraint

Take Home message:

- Data structures for calculating and storing problem solutions in a compressed form using an acyclic directed graph



MDD which contains three solutions :

(a,b) (a,a) (b,b)

4 IDEAS

- N-gram
- MDD to handle N-gram
- Integrate constraints in MDD
- Select best sentences

Idea 0

N-gram as an implicit constraint

Ngram : An implicit constraint

- Avoiding *generate & test*
- Taking account of meaning and grammar
- Reducing combinatorics

Idea (n-gram)

Step 1 : Corpus

Il est beau.

Mon amie est triste.

Elle est belle ce soir.

Step 2: set of 2-grams



Step 3: Chaining

Mon amie

→ amie est

est triste.

est beau.

est belle

→ belle ce

→ ce soir.

General Remark :

N of n-grams ↗

Quality of sentence ↗

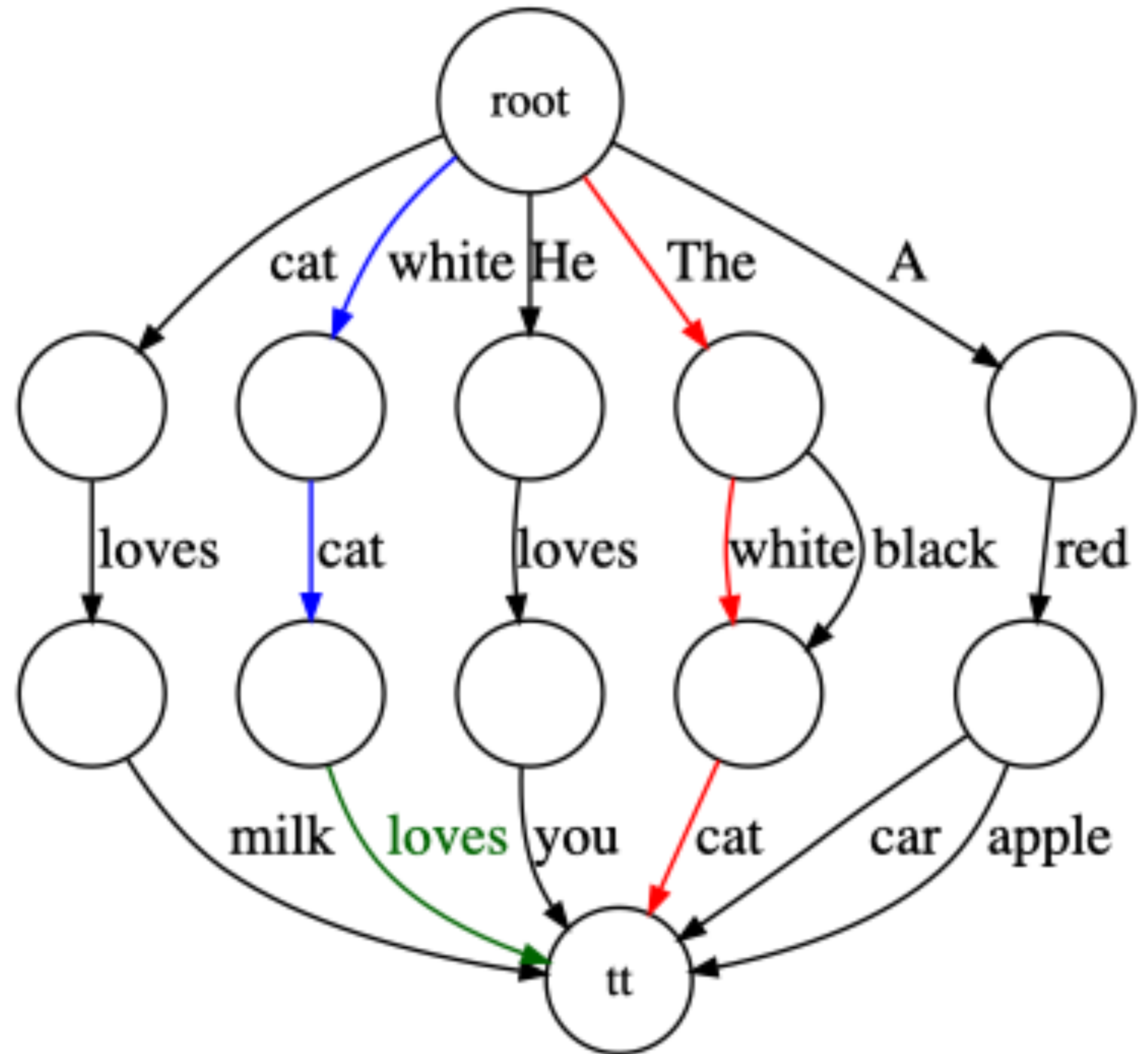
Number of sentence ↘

First idea

store and retrieve n-grams efficiently

Successions Constraint

- Assuming all n-grams are inserted in the MDD as solutions.
- MDD as a TRIE
- To store and reTRIEve n-grams.



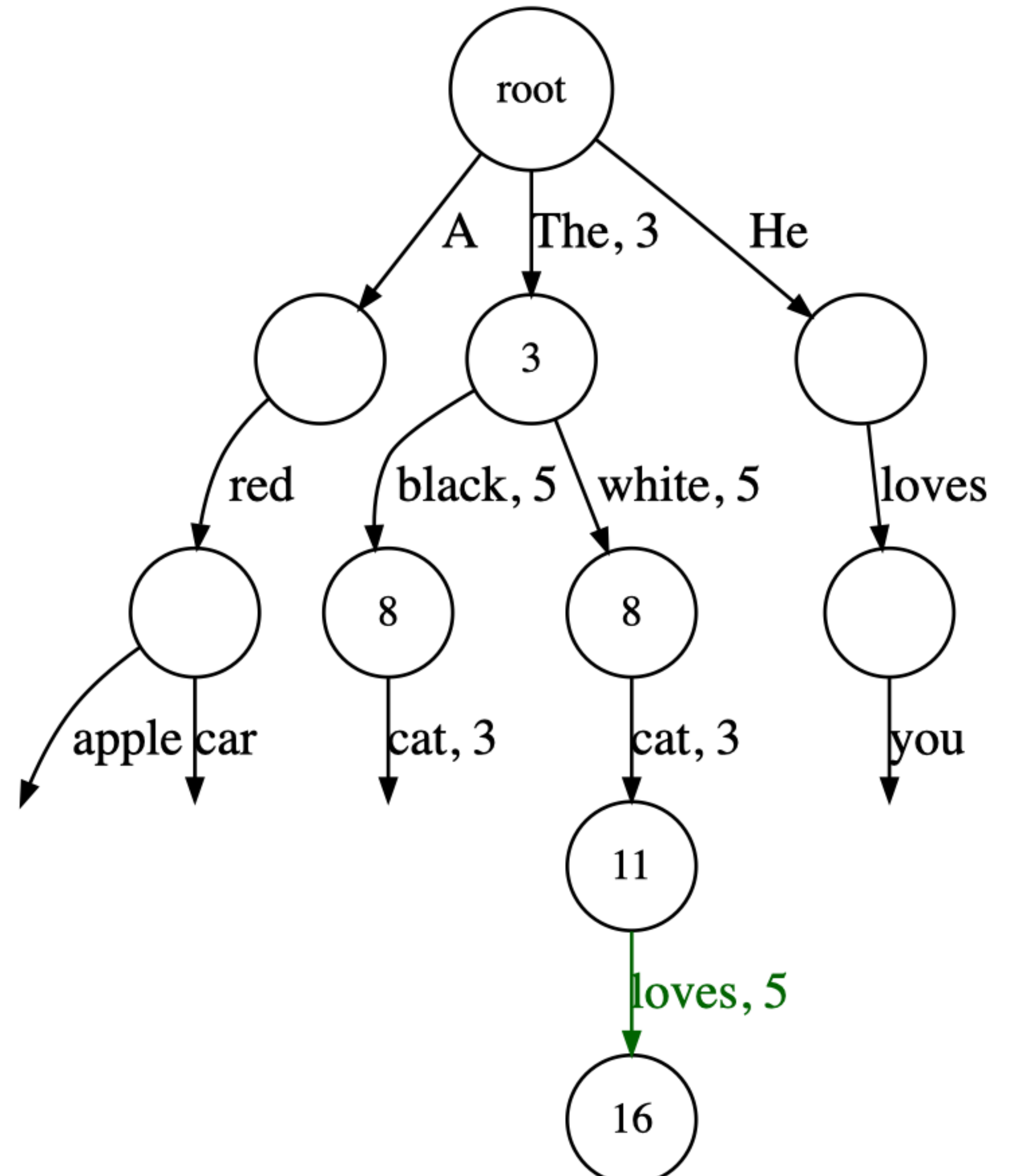
What is the next word of **The white cat** ?

Second idea

Integrate constraints on-the-fly

MDD Unfolding (top-down)

- Using the first MDD (successions)
- We compile the second one
 - Constraint are checked on-the-fly



Can we do better ?

Increase the n of n-gram !

● Je ne m'attends pas à ce que ses efforts soient récompensés

● Une femme sort de la maison et je n'ai pas le temps de fuir

● Une fille se tenait à quelques mètres de son pire cauchemar

● Ils vont dire à mon père que je n'ai pas envie de me marier

● Il ne comprend pas ce que tu es et tout ce que tu as décidé

● L'homme tourna la tête en direction de la voiture de police

● Son compagnon lui jeta un coup d'œil à travers les carreaux

● Une catastrophe est en train de m'aider à réaliser mon rêve

● Allez le dire à mon père que je n'ai pas envie de me marier

● Les images sont de plus en plus vers le cœur de la sorcière

- En 5-grammes, problème syntaxe et sens moins fréquent mais toujours présent

Third idea

Use an LLM to select best sentences

LLM sentences scoring : Perplexity

- Transformers (good for generation (chatGPT), but also to rank text):

Hugging Face is a startup based in New York City and Paris
p(word)

$$P(X_1 \dots X_n) = \prod_{i=1}^n P(X_i | X_1 \dots X_{i-1})$$

$$PPL(S_n) = \sqrt[n]{\frac{1}{P(w_1 w_2 w_3 \dots w_n)}}$$

- Perplexity is derived from Shannon entropy.
 - It quantify the uncertainty of a model with respect to a sample
 - Lower the better, range is [1 ; + inf[

Example of PPL ranking

PPL

Je ne m'attends pas à ce que ses efforts soient récompensés

9

Les images sont de plus en plus vers le cœur de la sorcière

60

Aucun de ses pieds nus sur les yeux de ce qui ne se passera

160

Il est tombé dans le vide avec une sorte de douceur absente

80

@

PROBLEM

STATE OF ARTS

OUR APPROACH

RESULT ★

CONCLUSION

Experimental conditions

- **Input** : 443 books belonging to the youth category (FR)
- **Input** : 75 books belonging to the fiction category (EN)
- **Evaluation** :
 - MNREAD **candidate** sentence set (syntax and meaning correct)
 - **Ineligible** set of sentences (syntax and/or meaning problems)
- **Software & Hardware** :
 - The model is implemented in Java 17 in an MDD solver (MDDLlib) @I3S.
 - The LLM use to rank sentences is GPT-2
 - Machine: Ubuntu 18.04 using an Intel(R) Xeon(R) Gold 5222 @ 3.80GHz CPU and 256 GB RAM.

Questions:

Are MNREAD sentences generated?

What is their quality ?

Are MNREAD sentences generated?

- YES ! In **5-grams** , with **443** books , we generate thousands of sentences (7028).

French Generated Sentences	PPL
Assise à la table de la salle à manger il y a un bon moment	6.23
Elle se pencha vers elle et lui donna un grand coup de pied	6.62
Et il y a des choses dont on ne peut pas dire la même chose	6.82
Nous avons à peine le temps de faire un tour dans la maison	6.83
Ses parents ne sont pas au courant de la maladie de sa mère	6.94
....	...
Il frappa à nouveau dans ses yeux et de voir la grue bouger	230
La poupée finit par tomber sur le sol et préparai mon bâton	241
Une hache reposait près de lui et de ne pas jouer aux héros	241
Quand le bout de la salle et frappa dans ses mains en coupe	242
Une hache reposait près de lui et de ne pas sortir ensemble	291



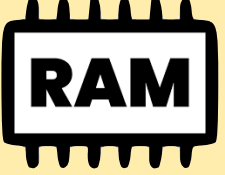

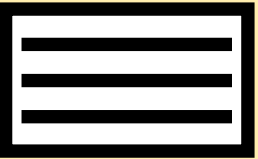
Are MNREAD sentences generated?

- YES ! In **5-grams** , with **75** books , we generate hundreds of sentences (204).

English Generated Sentences	PPL
And he had no idea what to do with the fact that she was in	17.9
It made me wonder if it was going to be able to do the work	20.1
You should be able to get out of bed to get out of the room	20.2
You need to get out of bed to get out of the room right now	21.2
No one will be able to get in and out of the room right now	21.4
....	...
The family in front of the double doors and into the branch	109
She paused outside the door to the back of my left shoulder	123
The difference here was now she had to deal with in my life	160
She hesitated at the door to the back of her nose and mouth	189
So strange to be on the edge of my chin between his fingers	224

Performances analysis

MNREAD sentence generation

					
FR	443	3Go	72s	7028	
EN	75	<<1Go	3s	204	

- Scoring takes roughly 1 hours for 7000 sentences. GPT-2 (lab server no GPU)
- Scoring takes roughly 30 mins for 7000 sentences. GPT-3 (OpenAI cloud)

PLAN

PROBLEM

STATE OF THE ART

OUR APPROACH

RESULT

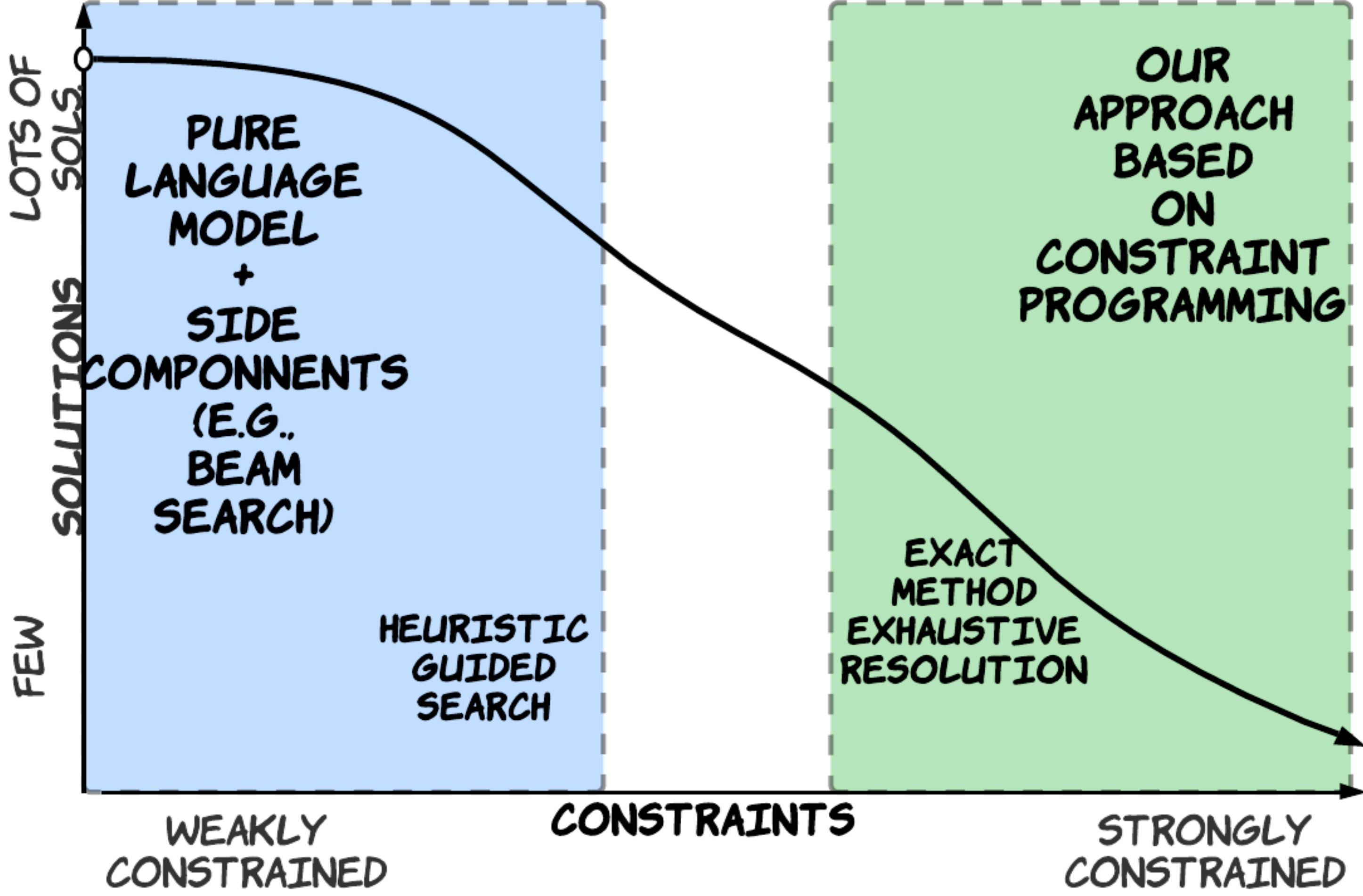
CONCLUSION ★

Conclusion

- Novel and creative approach that solves **constraints first** and then **select**
- Method: more suitable than generic methods for managing constraints (e.g., GPT, Bert) and more **flexible** than the ad-hoc method of Mansfield et al.
- Advantages: **modularity** (easy to add and/or remove rules), constraints taken into account at **generation stage**, applicable to **other languages**.

- Perspectives:
 - Constraint Programming - Machine Learning Bridge.

Overview



Thanks for your attention.



IJCAI/2023 MACAO

Paper:

Constraint First: A New MDD-based model to
generate sentences under constraints.

Aknowledgement:

Jean-Charles Régim, Université Côte Azur

Pierre Kornprobst, Inria

Aurélie Calabrese, Aix Marseille Université