KU LEUVEN

Extracting decision models from data and text

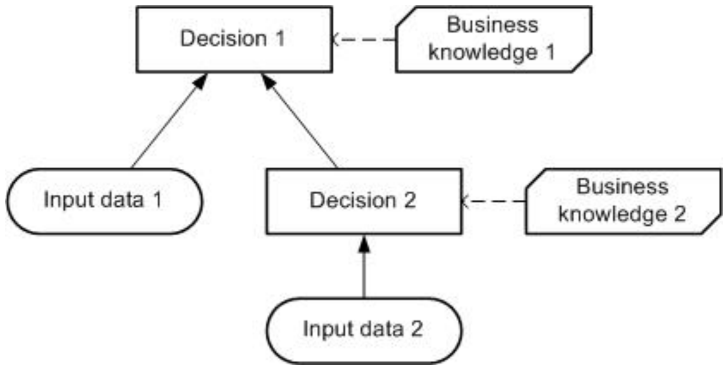
Jan Vanthienen, Vedavyas Etikala, Alexandre Goossens
LIRIS, Research Center for Management Informatics
KU Leuven, Belgium

RuleML Webinar, 2021

Overview

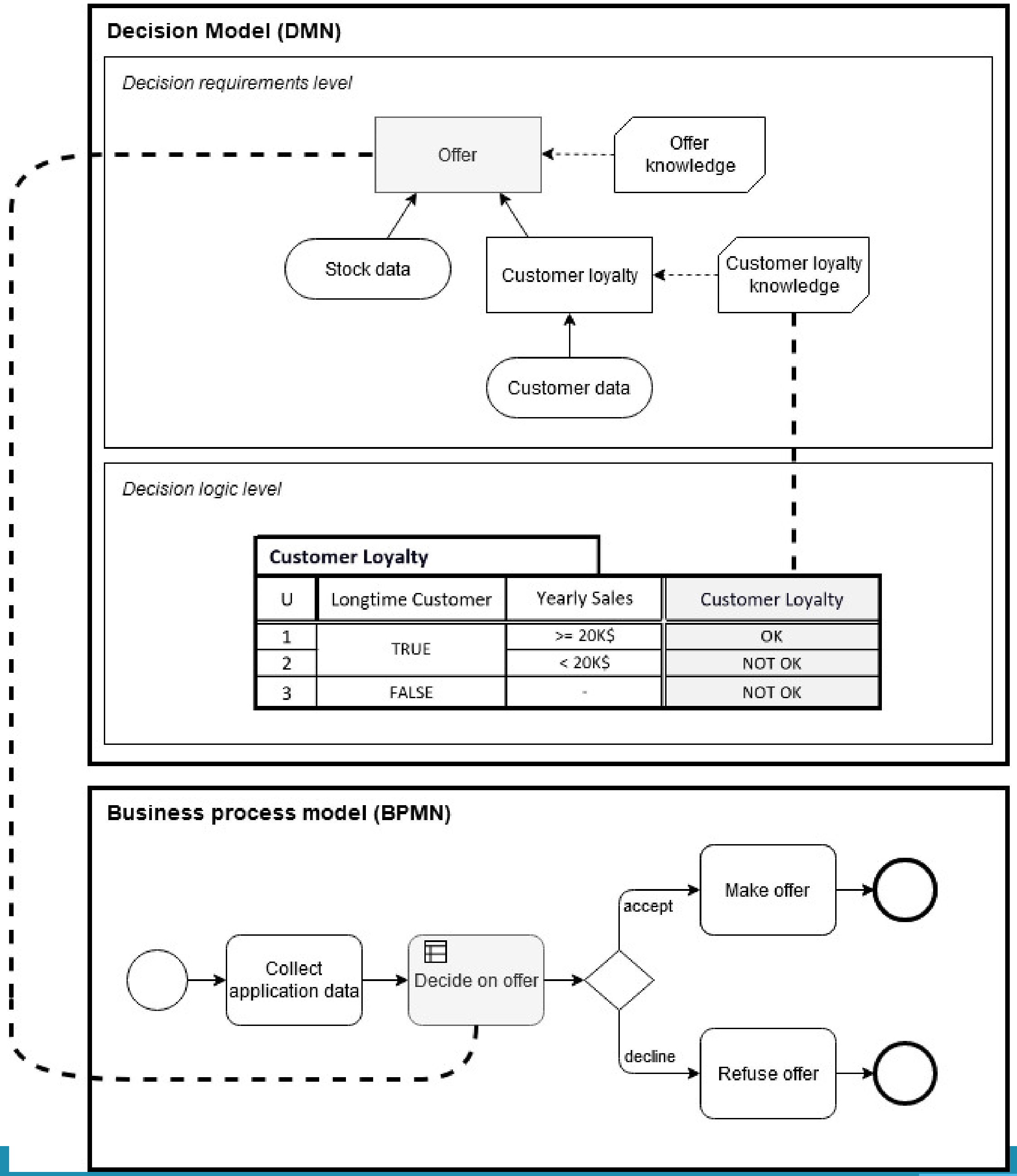
- Decision modeling
- Extracting decision models from data/cases
- Extracting decision models from text
- Challenges and future research
- Conclusion

Decision Modeling with DMN



Informaton requirement

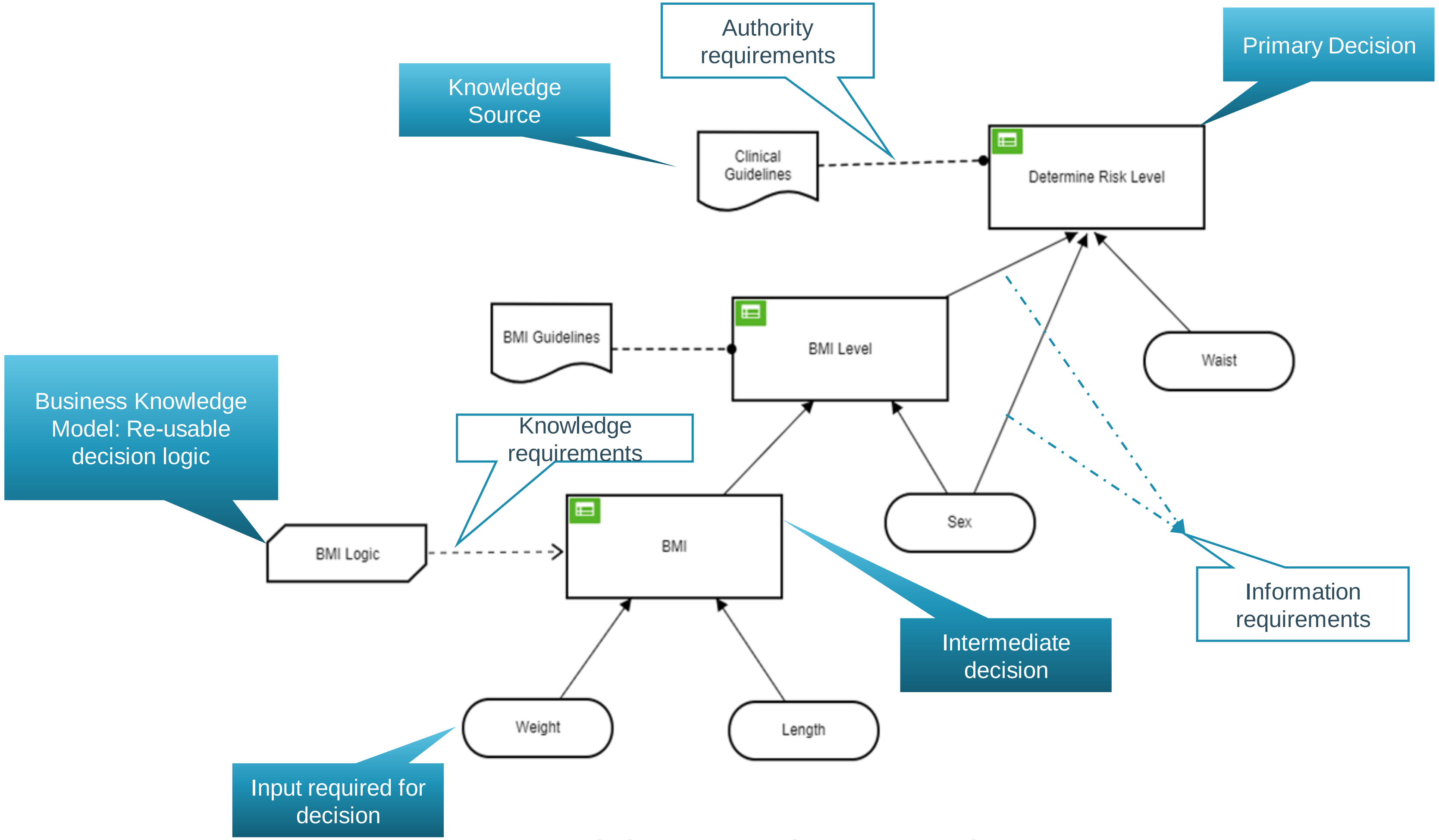
Knowledge requirement



Customer Loyalty

U	Longtime Customer	Yearly Sales	Customer Loyalty
1	TRUE	>= 20K$	OK
2		< 20K$	NOT OK
3	FALSE	-	NOT OK

Decision Requirements Diagram



Decision Requirement Diagram

Decision Logic

Primary Decision logic

Input conditions

Result

Determine Risk Level

RiskLevel

U	Input			Output
	BMI Level	Sex	Waist(cm)	Risk Level
	string	string	double	string
1	"Overweight"	"Male"	<=102	"Increased"
2	"Overweight"	"Male"	>102	"High"
3	"Overweight"	"Female"	<=88	"Increased"
4	"Overweight"	"Female"	>88	"High"
5	"Obese I"	"Male"	<= 102	"High"
6	"Obese I"	"Male"	>102	"Very High"
7	"Obese I"	"Female"	<=88	"High"
8	"Obese I"	"Female"	>88	"Very High"
9	"Obese II"	-	-	"Very High"

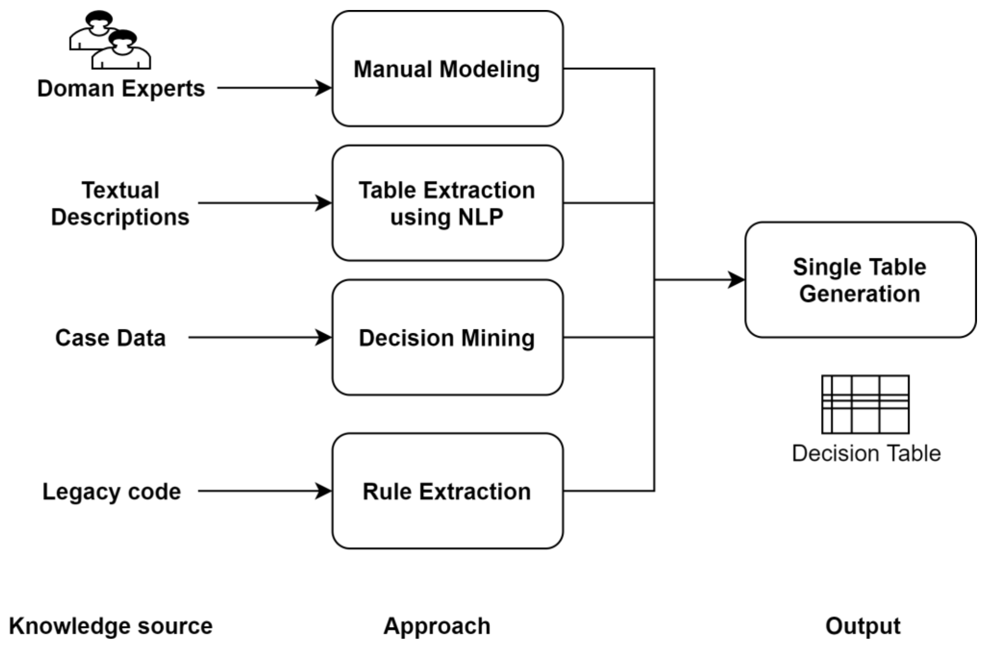
Each row is a decision rule

Decision Table

Decision rules and table extraction

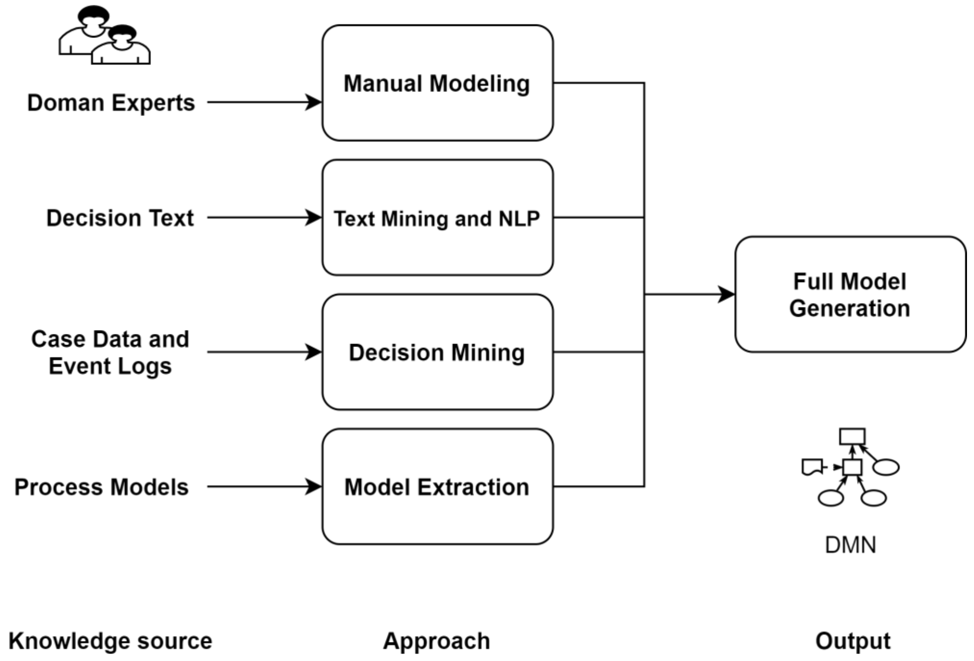
Existing approaches:

- Decision modeling methodology
- Extracting rules (and tables) from text
- Mining rules and tables from data
    (accuracy vs comprehensibility)
- Extracting rules from code



DRD (Decision requirements) extraction

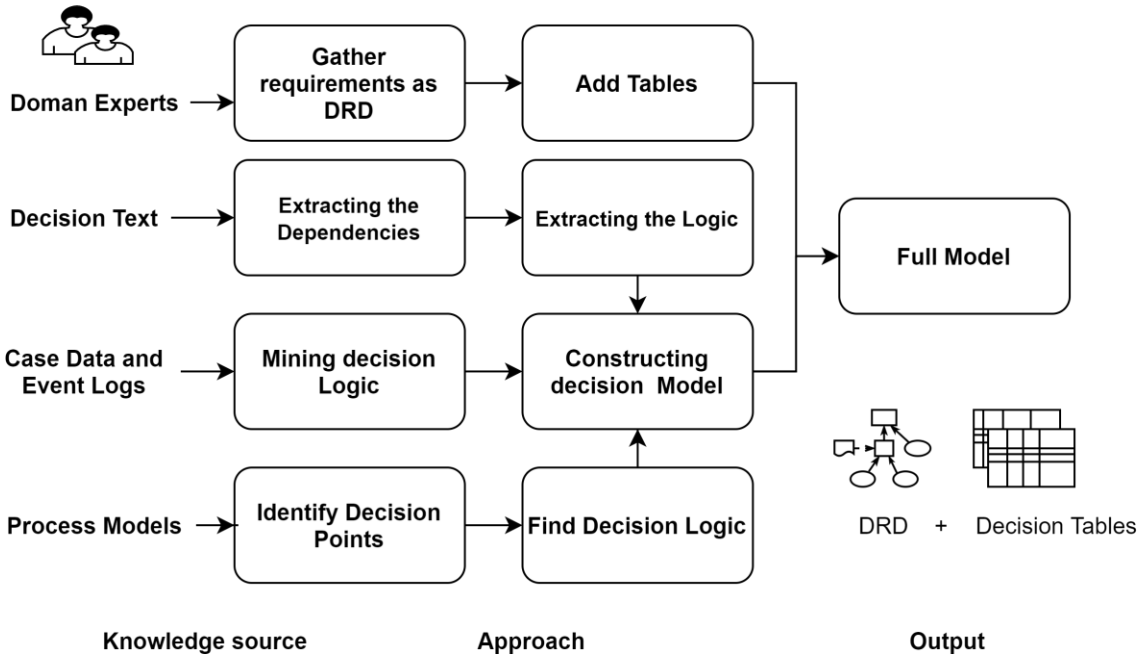
- Decision modeling methodology
- Extracting dependencies from text[1]
- Mining decisions and tables from data (process mining + data mining)
- Extracting DRD from process models



[1] *Etikala, V., Veldhoven, Z.V., Vanthienen, J.: Text2dec: Extracting decision dependencies from natural language text for automated DMN decision modelling. In: Business Process Management Workshops (2020)*

Full decision model extraction

- Decision modeling methodology
  (Vanthienen J., 1993 – Fish A., 2012 - Silver B., 2016)
- Extracting dependencies + rules from text
- Mining decisions models from data[1]
  (process mining + data mining)
- Extracting DMN from process models



[1] *De Smedt, J., Hasic, F., vanden Broucke, S., Vanthienen, J.: Holistic discovery of decision models from process execution data. Knowl. Based Syst. 183 (2019)*

Extracting decision models from data/cases

From data

- From case data to decision trees, rules or networks (analytics, rule learning)
- From case data to analytics models (ANN) and then decision table models (Baesens, Vanthienen et al., 2003)
- From case data to DMN DRD

From event logs and data

- From process event logs to process models (process mining)
- From process event logs + case data to process models + predictive models (decision mining) (e.g. Rozinat & van der Aalst, 2006)
- From process event logs + case data to integrated process & decision models (integrated mining) (e.g. De Smedt, Hasic, vanden Broucke & Vanthienen (2017).

Extracting decision models from text

An analogy: process model generation from natural language text (Friedrich et al.,2011)

- Deriving rules from text
- Deriving tables from text
- **Deriving DRDs and decision logic from text**

Three stages to extract decision models

1. **Which part of the text: classification**
2. Extracting dependencies
3. Extracting decision logic

Stage 1: Text Classification

Text sample: Medical Guidelines for Obesity

- *"The health risk level of a patient should be assessed from the obesity level, waist circumference and the sex of the patient. Furthermore, the degree of obesity should be determined from the BMI value and sex of the patient. Patient's height and weight are considered to calculate his BMI value."*
- *"When the patient's sex is a male and his BMI value is in between 25 and 29.9, then his obesity level is normal."*
- *"If patient' sex is female and BMI value is above 25.0 and less than 30, then obesity level is overweight . Where as, If BMI value is 30.0 or higher, obesity level falls within the obese I range."*

Declarative Sentences

Conditional Sentences

Concepts,
Dependencies
&
Logic

BMI Guidelines: https://www.nhlbi.nih.gov/files/docs/guidelines/ob_gdlns.pdf

*Clinical Guidelines on the Identification, Evaluation, and Treatment of Overweight and Obesity in Adults

Techniques considered for Sentence Classification

- **Deep Learning Classifiers**
  - **BERT for Sequence Classification**

    (Bidirectional Encoder Representations from Transformers)
    classify sentences into irrelevant, decision logic and decision dependency.
  - **Neural Network with GloVe as an Embedding Layer**

- **Non-Deep Learning Models**
  - multinomial logistic regression, Naive Bayes and support vector machines

Alexandre Goossens, Charlotte Parthoens, Michelle Claessens and Jan Vanthienen, Deep learning for the extraction of decision modelling components, In preparation, 2021.

Results

- The training set consists of 400 sentences and the test set contains 149 sentences.
  Both sets have a balanced distribution of the classes.
- BERT is able to retrieve all sentences labeled as dependency (Recall= 1.00) and is good at identifying the sentences labeled as logic (Recall = 0.86).

Table 2: Overview of results

Deep learning models					
Model	**Label**	**Precision**	**Recall**	**F1-score**	**Accuracy**
GloVe+MLP	Dependency	0.61	0.59	0.60	0.58
	Logic	0.56	0.63	0.59	
	Irrelevant	0.60	0.54	0.57	
GloVe + CNN	Dependency	0.74	0.50	0.60	0.65
	Logic	0.71	0.57	0.63	
	Irrelevant	0.60	0.82	.70	
BERT for sequence classification	Dependency	0.72	1.00	.84	**0.83**
	Logic	0.86	0.86	0.86	
	Irrelevant	0.91	0.70	0.79	
Non-deep learning models					
Model	**Label**	**Precision**	**Recall**	**F1-score**	**Accuracy**
BoW + Logistic Regression	Dependency	0.70	0.62	0.66	0.69
	Logic	0.81	0.57	0.67	
	Irrelevant	0.63	0.84	0.72	
BoW + Naïve Bayes	Dependency	0.66	0.85	0.74	**0.72**
	Logic	0.80	0.63	0.70	
	Irrelevant	0.71	0.72	0.71	
BoW + SVM	Dependency	0.67	0.59	0.62	0.66
	Logic	0.81	0.51	0.63	
	Irrelevant	0.60	0.84	0.70	
TF-IDF + Logistic Regression	Dependency	0.68	0.74	0.70	0.70
	Logic	0.93	0.53	0.67	
	Irrelevant	0.62	0.82	0.71	
TF-IDF + Naïve Bayes	Dependency	0.75	0.62	0.68	0.64
	Logic	0.70	0.45	0.55	
	Irrelevant	0.58	0.82	0.68	
TF-IDF + SVM	Dependency	0.65	0.82	0.73	0.71
	Logic	0.86	0.63	0.73	
	Irrelevant	0.66	0.72	0.69	

Stage 2: Extracting dependencies

1. Pattern based approach with NLP

2. Deep learning approach

Stage 2-1: A pattern based approach

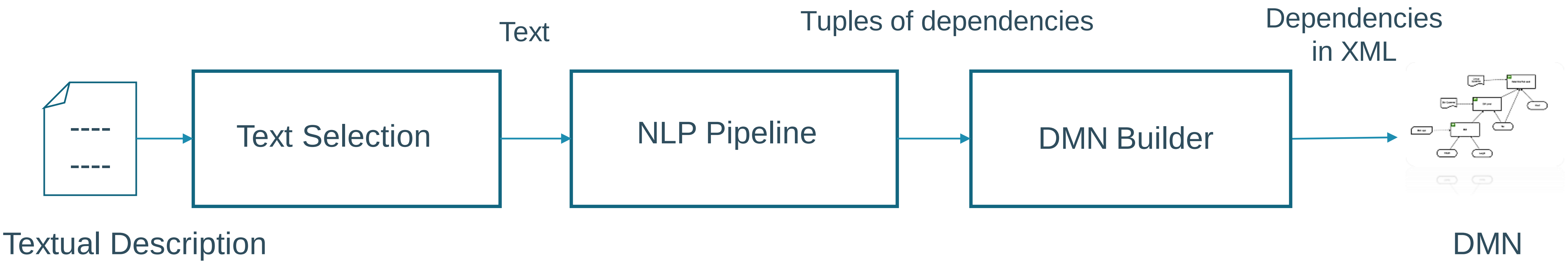


Fig: Three Stage Methodology of Tex2Dec

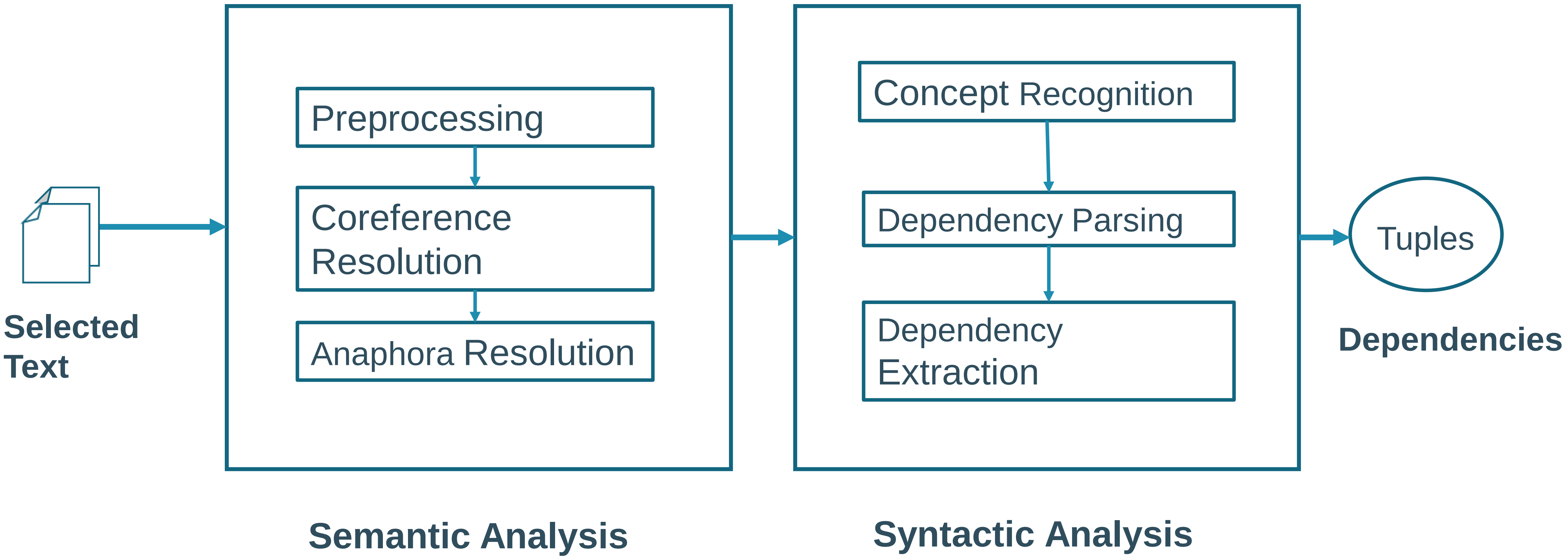
Vedavyas Etikala, Ziboud Van Veldhoven, Jan Vanthienen: Text2Dec: Extracting Decision Dependencies from Natural Language Text for Automated DMN Decision Modelling. Business Process Management Workshops 2020: 367-379

Sentence patterns

Dependency Pattern	Example	Base Concept (A)	Derived Concept (B)
Dec. Active A => B	*Patient's height determines his BMI value.*	*height*	*BMI value*
Dec. Passive B <= A	*Patient's BMI value is determined from his height.*	*height*	*BMI value*
Conditional A => B	*Unless the season is summer, do not plan a barbeque.*	*season*	*plan a barbeque*
Conditional B <= A	*A customer is loyal, if his annual sales are high.*	*annual sales*	*customer*

Patterns considered to extract dependencies.

Stage 2: NLP Pipeline.



Semantic Analysis

- Step A: Preprocessing

The health risk level of a patient should be assessed from the obesity level, waist circumference and the sex of the patient. Furthermore, the degree of obesity should be determined from the BMI value and sex of the patient. Patient's height and weight are considered to calculate his BMI value.

Text cleanup ↓

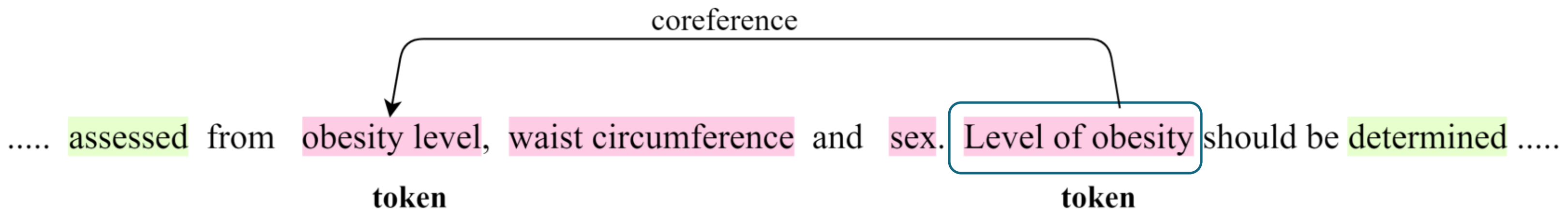
Remove determinants like ***the***, ***an*** and ***a***.
Remove non-inflicting prepositions or adverbs like ***futhermore***, ***thus***, ***but***.

- Result:

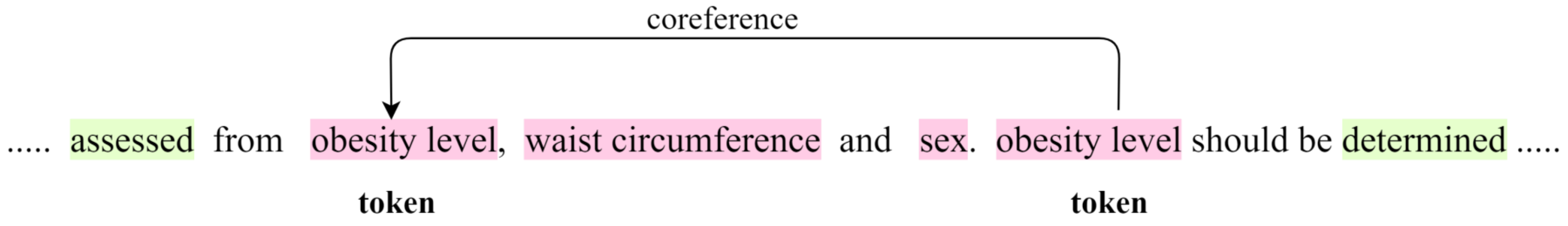
health risk level of a patient should be assessed from obesity level, waist circumference and sex of patient. degree of obesity should be determined from BMI value and sex of patient. Patient's height and weight are considered to calculate his BMI value.

Semantic Analysis

- Step B: Coreference resolution – fix cross referred concepts

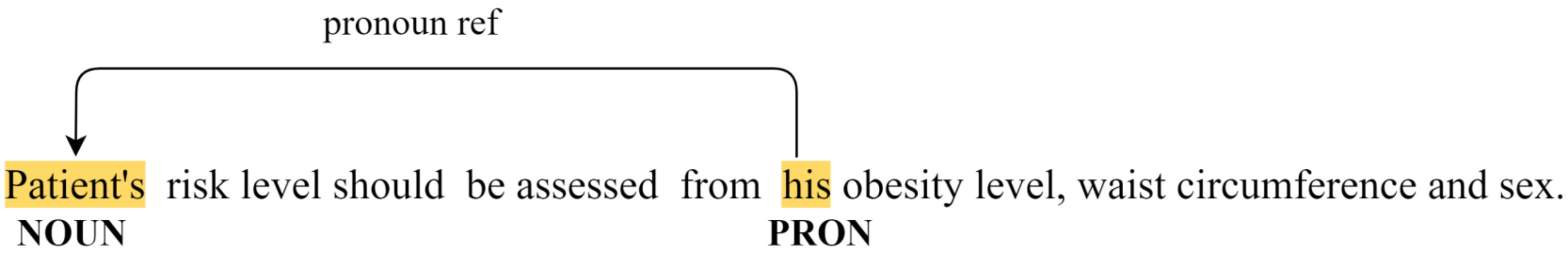


- Result :

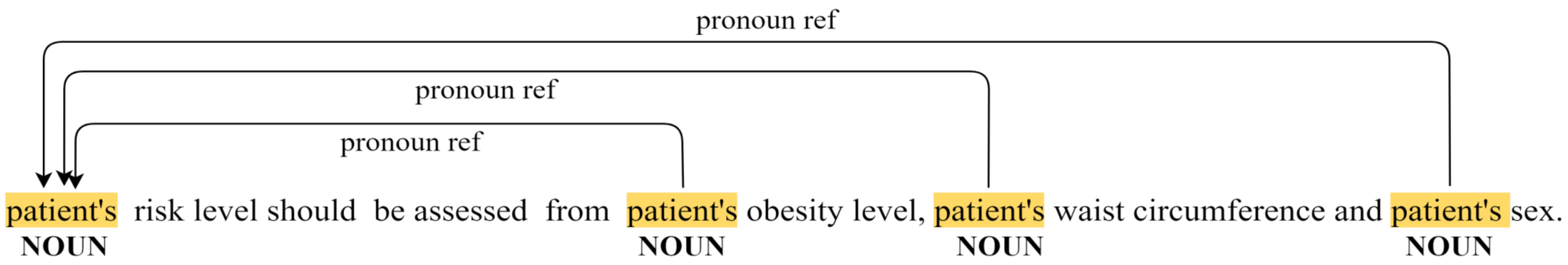


Semantic Analysis

- Step C: Anaphora resolution – Fix pronoun references and ownerships



- Result :



Syntactic Analysis

- Step D: Concept Recognition – Identifing nouns and noun phrases

noun phrase | noun phrase | noun phrase

health risk level should be assessed from obesity level, waist circumference and sex

NOUN NOUN NOUN NOUN NOUN NOUN NOUN NOUN

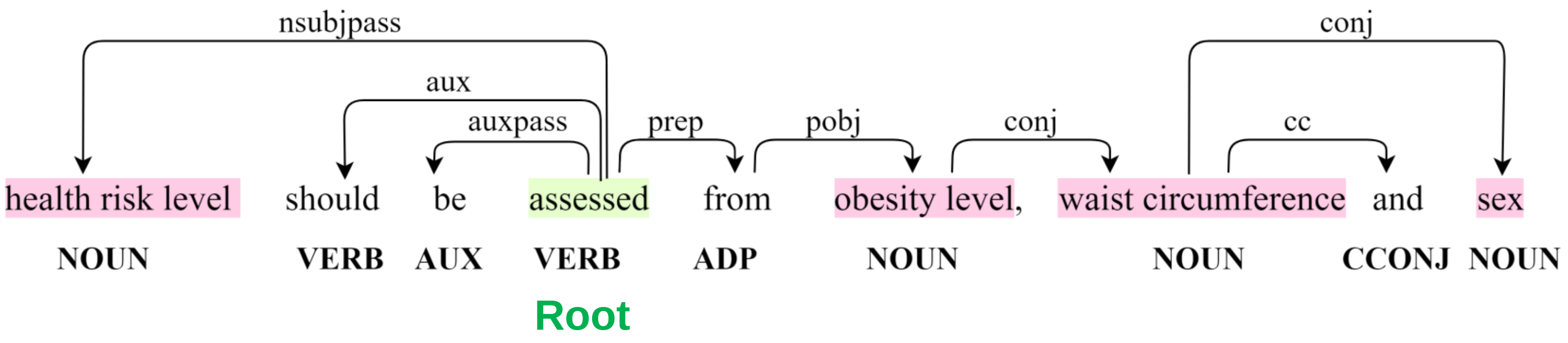
- Result :

health risk level should be assessed from obesity level, waist circumference and sex

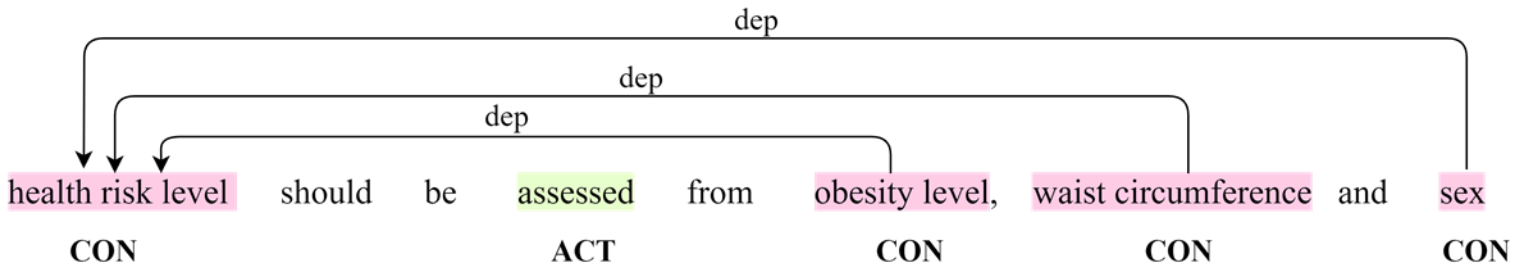
CON CON CON CON

Syntactic Analysis

- Step E: Dependency parsing – Identify the root action verb



- Step F: Dependency extraction based on action verbs



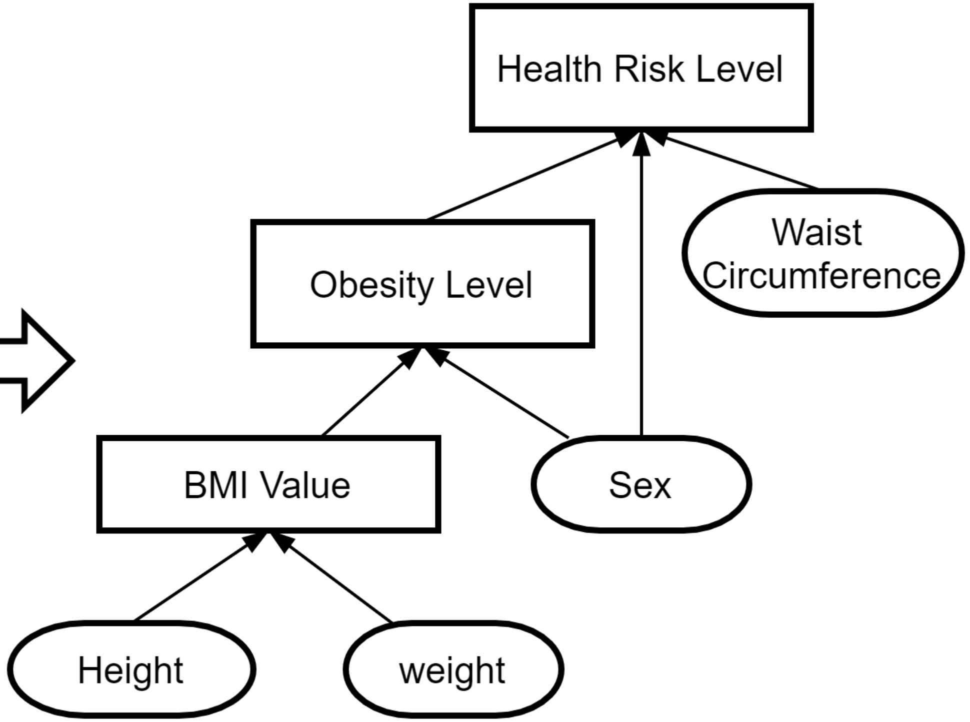
Example

S1: *The **health risk level** of a patient should be assessed from the **obesity level**, **waist circumference** and the **sex of the patient**.*

S2: *Furthermore, the **degree of obesity** should be determined from the **BMI value** and **sex of the patient**.*

S3: *Patient's **height** and **weight** are considered to calculate his **BMI value**.*

S4: *If the **weight of the patient** given in kgs and **height of patient** given in meters, then the **BMI value** is weight/(height*height).*



Stage 2-2: a deep learning approach

- Using inside-outside-beginning (IOB) tagging format for base concepts, derived concepts and action verbs

- **Two techniques were investigated to extract tags**
  - **BERT**

    (Bidirectional Encoder Representations from Transformers)
    classify sentences into irrelevant, decision logic and decision dependency.
  - **Bi-LSTM-CRF**

    (Bi-directional-Long Short-Term Memory- Conditional Random Field

Alexandre Goossens, Charlotte Parthoens, Michelle Claessens and Jan Vanthienen, Extracting decision dependencies and conditional clauses using deep learning, In preparation, 2021.

Results

- The training set consists of 195 explicit and 245 conditional dependency sentences, manually tagged.
- The test set contains 60 and 82 sentences.

Dependency Extraction		BERT base-uncased without stopword removal			BI-LSTM-CRF without preprocessing		
Explicit Dependency sentences	**Level**	**Precision**	**Recall**	**F1-score**	**Precision**	**Recall**	**F1-score**
	B-DER	0.79 ± 0.02	0.87 ± 0.03	0.83 ± 0.02	0.61 ± 0.06	0.62 ± 0.03	0.62 ± 0.02
	I-DER	0.84 ± 0.05	0.86 ± 0.03	0.85 ± 0.02	0.78 ± 0.02	0.63± 0.06	0.70 ± 0.03
	B-BAS	0.79 ± 0.02	0.88 ± 0.01	0.83 ± 0.01	0.58 ± 0.03	0.83± 0.02	0.68 ± 0.02
	I-BAS	0.87 ± 0.01	0.86 ± 0.05	0.86 ± 0.02	0.68 ± 0.02	0.71± 0.02	0.70 ± 0.01
	B-ACT	0.80 ± 0.02	0.93 ± 0.01	0.86 ± 0.01	0.82 ± 0.01	0.79 ± 0.00	0.80 ± 0.01
	AVG_MICRO	0.83 ± 0.02	0.87 ± 0.02	0.85 ± 0.02	0.68 ± 0.01	0.71 ± 0.01	0.70 ± 0.01
	AVG_MACRO	**0.82 ± 0.01**	**0.88 ± 0.02**	**0.85 ± 0.01**	0.70 ± 0.01	0.72 ± 0.01	0.70 ± 0.01
Conditional sentences	**B-DER**	0.89 ± 0.02	0.96 ± 0.02	0.92 ± 0.01	0.70 ± 0.08	0.78 ± 0.05	0.73 ± 0.03
	I-DER	0.90 ± 0.03	0.90 ± 0.03	0.90 ± 0.01	0.67 ± 0.11	0.72 ± 0.04	0.69 ± 0.05
	B-BAS	0.76 ± 0.01	0.85 ± 0.03	0.80 ± 0.01	0.55 ± 0.06	0.62 ± 0.07	0.57 ± 0.04
	I-BAS	0.93 ± 0.02	0.70 ± 0.02	0.80 ± 0.01	0.80 ± 0.04	0.33 ± 0.10	0.47 ± 0.10
	AVG_MICRO	0.88 ± 0.01	0.83 ± 0.01	0.85 ± 0.01	0.66 ± 0.07	0.58 ± 0.03	0.61 ± 0.05
	AVG_MACRO	**0.87 ± 0.01**	**0.85 ± 0.01**	**0.86 ± 0.01**	0.68 ± 0.05	0.62 ± 0.02	0.61 ± 0.05

Fig. 3: Results for dependency tag extraction

Stage 3: extracting decision logic

- "When the patient's sex is a male and his BMI value is in between 25 and 29.9, then his obesity level is normal."

Condition part | Consequence

When the patient's sex is a male and his BMI value is in between 25 and 29.9, then his obesity level is normal.

VAR VAL VAR VAL VAR VAL

Extracted Rule: **IF** BMI value in [25, 29.9] **AND** sex = male **THEN** obesity level = normal

3-1: Sentence patterns

Sentence Pattern	Example	Condition	Consequence
Explicit IF - THEN	*If patient' BMI value is above 25.0 and less than 30, then obesity level is overweight*	*BMI value in [25.0, 30]*	*obesity level = overweight*
Synonym IF- THEN	*Unless the season is summer, do not plan a barbeque.*	*Season = summer*	*plan a barbecue = true*
Implicit IF-THEN	*Any customer with high annual sales is loyal.*	*annual sales = high*	*customer = loyal*

Patterns considered to extract logical rules.

3-2: Deep learning approach

- The training set consists of 264 conditional sentences, manually tagged.
- The test set contains 82 sentences.
- Separate condition part and consequence part.

Logic Extraction		BERT base-uncased without stopword removal			BI-LSTM-CRF without preprocessing		
Conditional sentences	Level	Precision	Recall	F1-Score	Precision	Recall	F1-Score
	B-CONS	0.88 ± 0.01	0.89 ± 0.02	0.88 ± 0.00	0.67 ± 0.07	0.53 ± 0.04	0.59 ± 0.04
	I-CONS	0.91 ± 0.01	0.94 ± 0.02	0.93 ± 0.02	0.82 ± 0.02	0.62 ± 0.09	0.70 ± 0.06
	B-COND	0.87 ± 0.00	0.94 ± 0.01	0.90 ± 0.00	0.71 ± 0.04	0.73 ± 0.06	0.71 ± 0.03
	I-COND	0.93 ± 0.03	0.89 ± 0.01	0.91 ± 0.02	0.73 ± 0.05	0.77 ± 0.03	0.75 ± 0.02
	B-ELSE	0.91 ± 0.00	0.87 ± 0.05	0.89 ± 0.03	0.36 ± 0.17	0.71 ± 0.08	0.47 ± 0.15
	I-ELSE	0.95 ± 0.05	0.97 ± 0.02	0.96 ± 0.01	0.31 ± 0.11	0.93 ± 0.03	0.46 ± 0.11
	B-EXCE	1.00 ± 0.00	0.80 ± 0.00	0.89 ± 0.00	1.00 ± 0.00	0.40 ± 0.14	0.56 ± 0.15
	I-EXCE	1.00 ± 0.00	0.65 ± 0.14	0.78 ± 0.10	1.00 ± 0.00	0.40 ± 0.12	0.56 ± 0.13
	AVG_MICRO	0.92 ± 0.02	0.92 ± 0.02	0.92 ± 0.02	0.69 ± 0.03	0.69 ± 0.03	0.69 ± 0.03
	AVG_MACRO	**0.93 ± 0.00**	**0.87 ± 0.00**	**0.89 ± 0.00**	0.70 ± 0.03	0.66 ± 0.05	0.60 ± 0.05

Fig. 4: Results for logic tag extraction

Alexandre Goossens, Charlotte Parthoens, Michelle Claessens and Jan Vanthienen, Extracting decision dependencies and conditional clauses using deep learning, In preparation, 2021.

Early results

Eligibility for smallpox vaccine is depended on risk and outbreak.

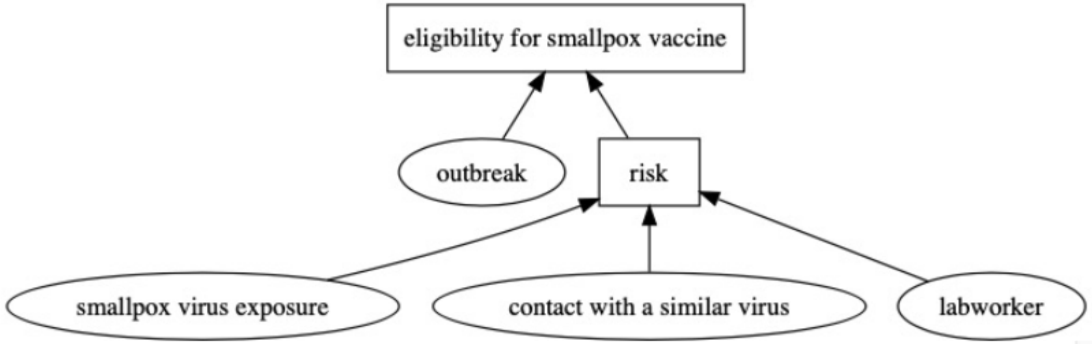
If you are at risk or there is an outbreak then eligibility for smallpox vaccine is true.

Risk is determined from contact with a similar virus, labworker or smallpox virus exposure.

You are at risk if contact with a similar virus is true and you are a labworker.
You are also at risk if there is smallpox virus exposure.

Conditional clause	Exception clause	Consequence clause	Else clause
you are at risk or there is outbreak	/	eligibility for smallpox vaccine is true	/
contact with similar virus is true and you are labworker	/	you are at risk	/
there is smallpox virus exposure	/	you are also at risk	/

LIST OF CONCEPTS =
outbreak
smallpox virus exposure
eligibility for smallpox vaccine
contact with a similar virus
risk
labworker
contact with a virus



Your set of dependencies can further help you to identify incorrect concepts.

LIST OF DEPENDENCIES =
('determined', 'labworker', 'risk')
('depended', 'risk', 'eligibility for smallpox vaccine')
('determined', 'contact with a similar virus', 'risk')
('determined', 'smallpox virus exposure', 'risk')
('depended', 'outbreak', 'eligibility for smallpox vaccine')
('None', 'contact with a virus', 'risk')

Challenges and Future research

- Linguistic Challenges from NLP
  - Ambiguities
  - Incompleteness
- Full automation of model extraction requires:
  - Understanding concepts and values
  - Order of the rules
  - Table hit policies, decompositions
- Huge potential for further study
  - Comparing pattern based approaches and deep learning
  - Digital Automation with DMN

Conclusion

- While data science and data analytics are doing just fine on their own, integrating it with DMN can not only add ***explainability*** but actually improve ***accuracy***.
- Knowledge based systems rely mostly on textual guidelines, policies or regulations as their knowledge sources. Automatic extraction provides an insight about how a textual resource (e.g. a clinical guideline document) could be made **interpretable** not only to domain experts but also to computers systems. The same document/model to applications and users.
- Cuts down the modeling **time** notably.

Some references

- Arco, L., Napoles, G., Vanhoenshoven, F., Lara, A.L., Casas, G., Vanhoof, K.: Natural language techniques supporting decision modelers. Data Mining and Knowledge Discovery 35(1), 290{320 (2021)
- De Smedt, J., Hasić, F., Vanden Broucke, S., Vanthienen, J.: Towards a Holistic Discovery of Decisions in Process-Aware Information Systems, presented at BPM 2017.
- Dragoni, M., Villata, S., Rizzi, W., Governatori, G.: Combining NLP approaches for rule extraction from legal documents. In: 1st Workshop on MIning and Reasoning with Legal texts (MIREL 2016) (2016)
- Friedrich, F., Mendling, J., Puhlmann, F.: Process model generation from natural language text. In: Advanced Information Systems Engineering. pp. 482–496. Springer (2011)
- Janssens, L., Bazhenova, E., Smedt, J.D., Vanthienen, J., Denecker, M.: Consistent integration of decision (DMN) and process (BPMN) models. In: CAiSE Forum. Volume 1612 of CEUR Workshop Proceedings., CEUR-WS.org (2016) 121–128.
- Marneffe, M.-C. D. and Manning, C. D. (2010). Stanford typed dependencies manual. 20090110 Httpnlp Stanford, 40(September):1–22.
- Silver, B. (2016). DMN Method & Style The practitioner's guide to decision modeling with business rules. Cody-Cassidy Press, Altadena, CA.
- Vedavyas Etikala, Ziboud Van Veldhoven, Jan Vanthienen: Text2Dec: Extracting Decision Dependencies from Natural Language Text for Automated DMN Decision Modelling. Business Process Management Workshops 2020: 367-379

Thank you ☺