KLU
KÜHNE LOGISTICS UNIVERSITY

NATURAL LANGUAGE PROCESSING IN PROCESS ANALYTICS

How Textual Data Can Help to Analyze and Improve Organizations

Prof. Dr. Henrik Leopold
Kühne Logistics University

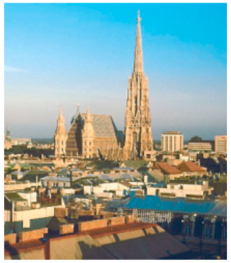
Technische Universität Dortmund
Monday, November 15, 2021

ABOUT MYSELF

BA Berlin
Information Systems (B.Sc.)
2005 – 2008

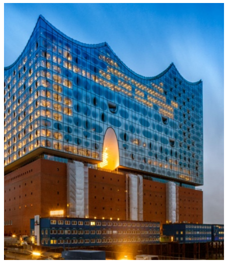
Humboldt University Berlin
Information Systems (M.Sc.)
2008 – 2010

Humboldt University Berlin / UNIRIO
Business Process Analytics (PhD)
2011 – 2013



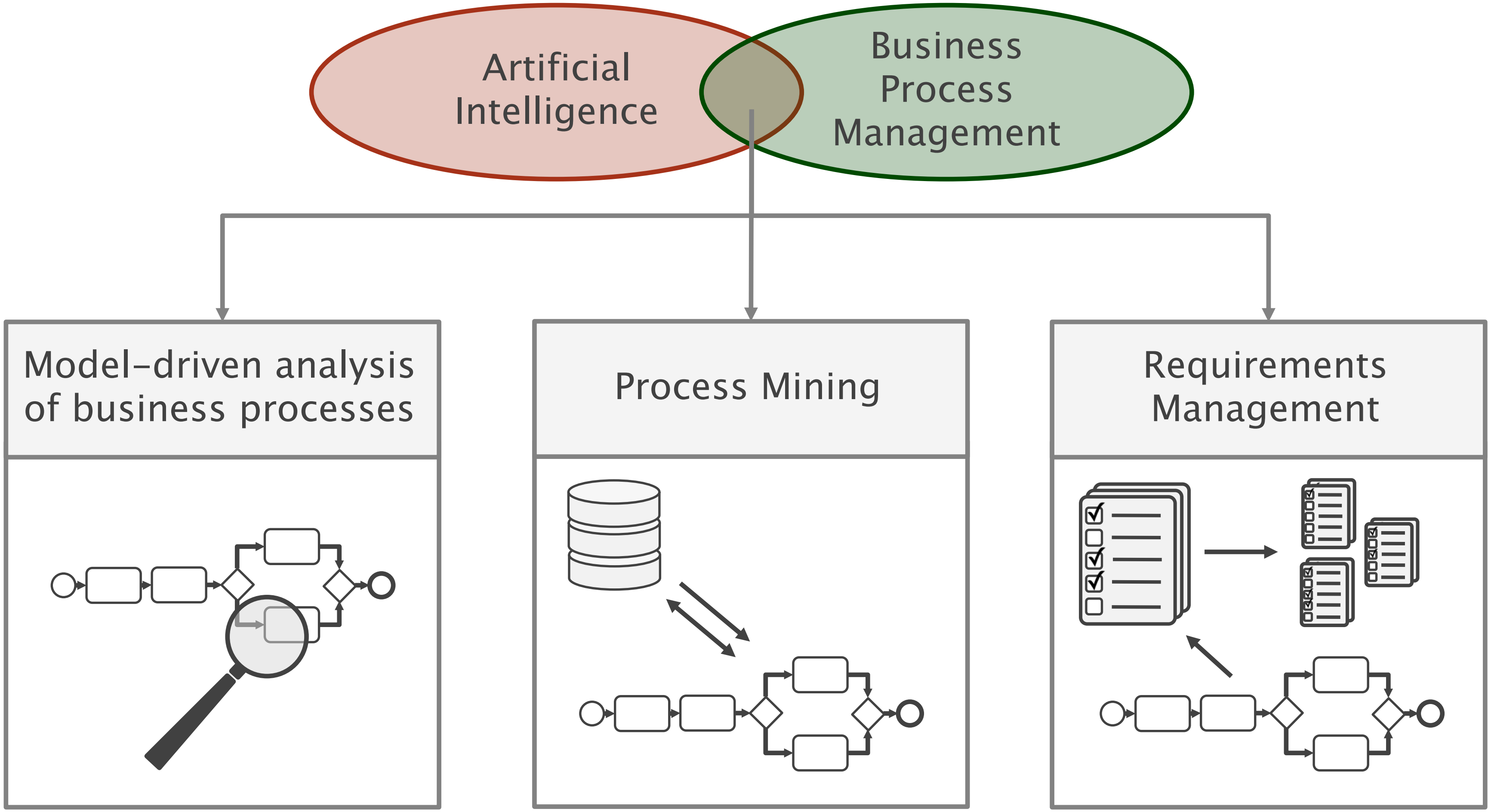
WU Vienna
Post-Doc
2014 – 2015

Vrije Universiteit Amsterdam
Assistant Professor
2015 – 2019



KLU / HPI
Associate Professor
2019 – today

ABOUT MY RESEARCH



Artificial Intelligence

Business Process Management

Model-driven analysis of business processes

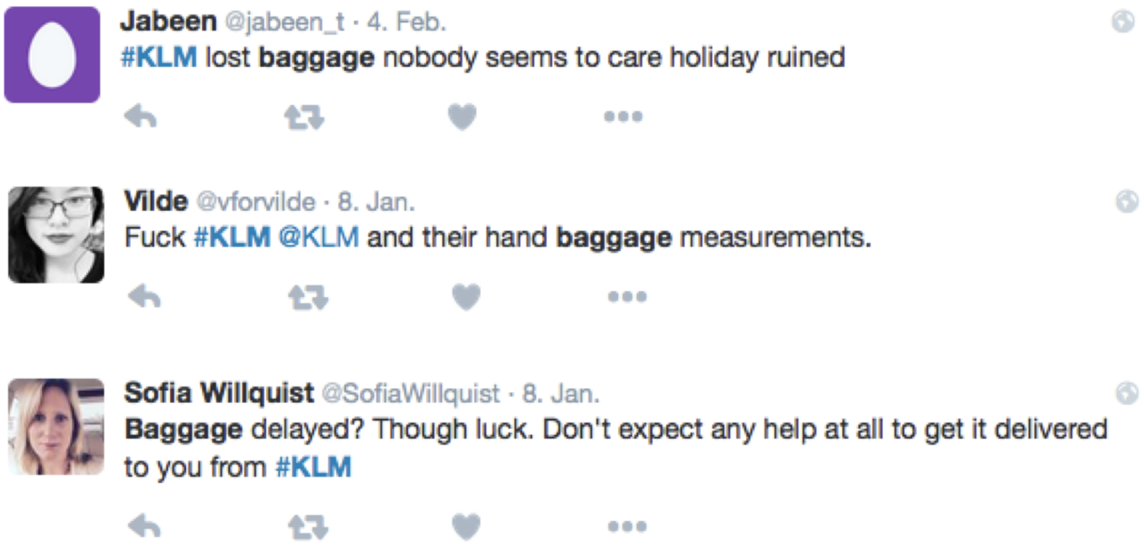
Process Mining

Requirements Management

AGENDA

1. Why Natural Language Processing?
2. Background on Linguistics
3. Technology for Analyzing Natural Language Documents
4. NLP in Process Models
5. NLP in Textual Process Descriptions
6. NLP in Event Logs

WHAT CAN WE LEARN FROM THIS?



Jabeen @jabeen_t · 4. Feb.
#KLM lost **baggage** nobody seems to care holiday ruined

Vilde @vforvilde · 8. Jan.
Fuck #KLM @KLM and their hand **baggage** measurements.

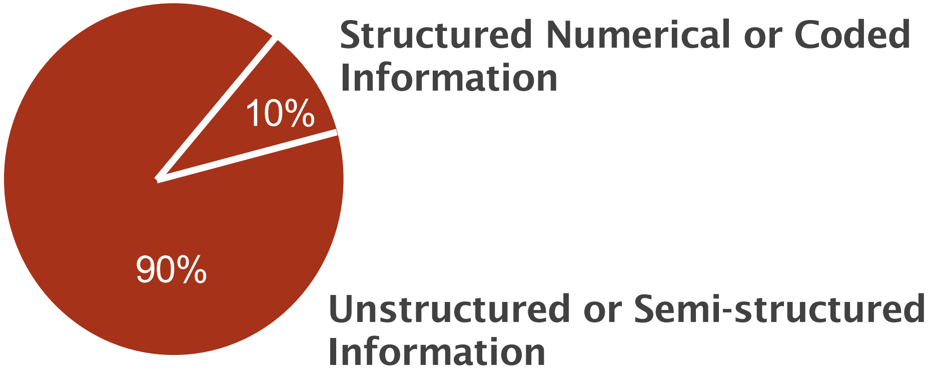
Sofia Willquist @SofiaWillquist · 8. Jan.
Baggage delayed? Though luck. Don't expect any help at all to get it delivered to you from #KLM

WHAT CAN WE LEARN FROM THIS?

At the beginning the customer perceives that her subscribed service has degraded. A list with all the problem parameters is then sent to the Customer Service department of TELECO. At the customer service an employee enters (based on the received data) a problem report into system T. Then the problem report is compared to the customer SLA to identify what the extent and the details of the service degradation are. Based on this, the necessary counter measures are determined including their respective priorities. An electronic service then determines the significance of the customer based on information that has been collected during the history of the contractual relationship. In case the customer is premium, the process will link to an extra problem fix process (this process will not be detailed here). In case the customer is of certain significance which would affect the counter measures previously decided upon, the process goes back to re-prioritize these measures otherwise the process continues. Taking together the information (i.e. contract commitment data + prioritized actions) a detailed problem report is created. The detailed problem report is then sent to Service Management. Service Management deals on a first level with violations of quality in services that are provided to customers. After receiving the detailed problem report, Service management investigates whether the problem is analyzable at the level of their department or whether the problem may be located at Resource Provisioning. In case Service Management assesses the problem to be not analyzable by themselves, the detailed problem report is sent out to Resource Provisioning. If Service Management is sure they can analyze it, they perform the analysis and based on the outcome they create a trouble report that indicates the type of problem. After Resource Provisioning receives the detailed problem report, it is checked whether there are any possible problems. If no problems are detected, a notification about the normal service execution is created. If a problem is detected this will be analyzed by Resource Provisioning and a trouble report is created. Either trouble report or the normal execution notification will be included in a status report and sent back to Service Management. Service Management then prepares the final status report based on the received information. Subsequently it has to be determined what counter measures should be taken depending on the information in the final status report. Three alternative process paths may be taken. For the case that no problem was detected at all, the actual service performance is sent back to the Customer Service. For the case that minor corrective actions are required, Service Management will undertake corrective actions by themselves. Subsequently, the problem resolution report is created and then sent out to Customer Service. After sending, this process path of Service Management ends. For the case that automatic resource restoration from Resource Provisioning is required, Service Management must create a request for automatic resource restoration. This message is then sent to Resource Provisioning. Resource Provisioning has been on-hold and waiting for a restoration request but this must happen within 2 days after the status report was sent out, otherwise Resource Provisioning terminates the process. After the restoration request is received, all possible errors are tracked. Based on the tracked errors, all necessary corrective actions are undertaken by Resource Provisioning. Then a trouble-shooting report is created. This report is sent out to Service Management; then the process ends. The trouble-shooting report is received by Service Management and this information goes then into the creation of the problem resolution report just as described for ii). Customer Service either receives the actual service performance (if there was no problem) or the problem resolution report. Then, two concurrent activities are triggered, i.e. i) a report is created for the customer which details the current service performance and the resolution of the problem, and ii) an SLA violation rebate is reported to Billing \& Collections who will adjust the billing. The report for the customer is sent out to her. After all three activities are completed the process ends within Customer Service. After the customer then receives the report about service performance and problem resolution from Customer Service, the process flow at the customer also ends.

RELEVANCE OF NLP

- Approximately **90%** of the world's data is held in unstructured formats
- Information intensive business processes demand that we transcend from simple document retrieval to "knowledge" discovery



POTENTIAL OF NATURAL LANGUAGE PROCESSING

- Textual descriptions are valuable resource for process analysis
- They include information about:
  - How the process really goes
  - Problems that occur during its execution
  - How customers perceive the provided services
  - How to improve the process
  - …

How to make us of this in an automated fashion?

AGENDA

1. Why Natural Language Processing?
2. Background on Linguistics
3. Technology for Analyzing Natural Language Documents
4. NLP in Process Models
5. NLP in Textual Process Descriptions
6. NLP in Event Logs

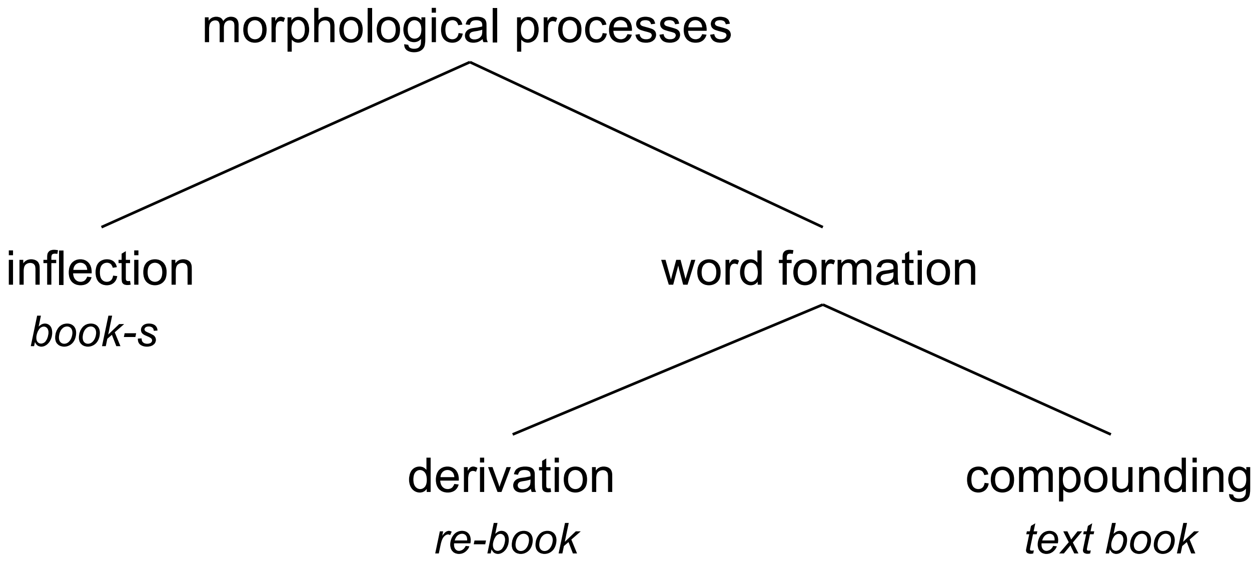
LEVELS OF LANGUAGE ANALYSIS

1. Phonology
   - Study of sound systems of languages
2. Morphology
   - Study of structure of words
3. Syntax
   - Study of organization of words in sentences
4. Semantics
   - Study of meaning in language
5. Pragmatics
   - Study of language in use
6. Discourse
   - Study of language in a particular context

MORPHOLOGY

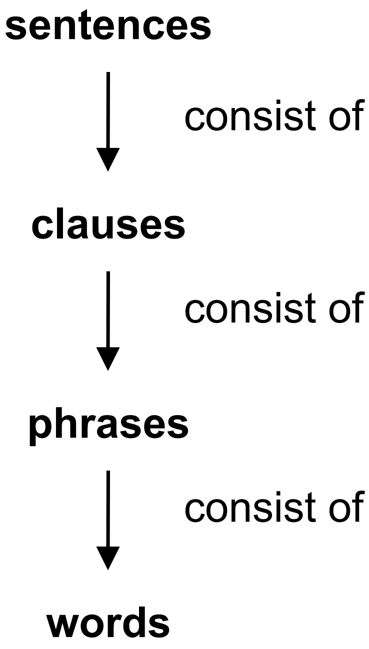
- A morpheme is the smallest meaning-bearing unit of a language
- Two broad classes of morphemes:
  - **Stems**: “main” morpheme of the word, supplying meaning
  - **Affixes**: bits and pieces that combine with stems to modify their meanings and grammatical functions (prefixes, suffixes, circumfixes, infixes)
    - Unlike
    - Trying
    - Unreadable

MORPHOLOGICAL PROCESSES



SYNTAX

- Syntax is concerned with how words can be combined to phrases, clauses, and sentences



sentences

consist of

clauses

consist of

phrases

consist of

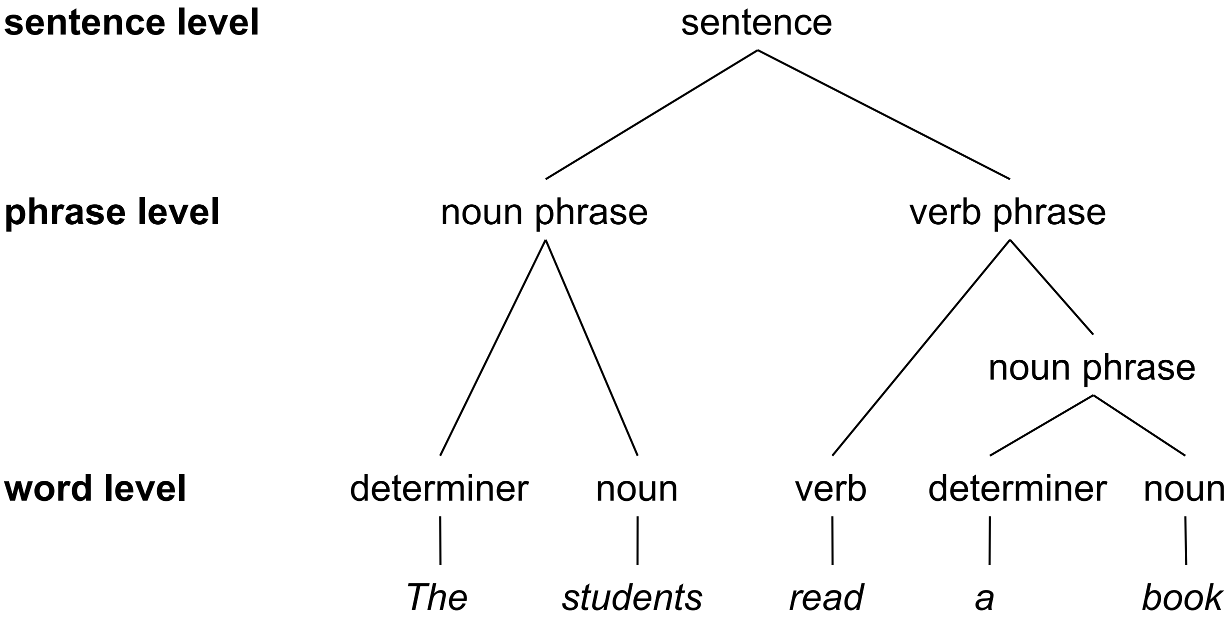
words

PARTS OF SPEECH

A **part of speech** is a category of words with similar grammatical properties:

1. **Noun**: people, animals, concepts, things (e.g. “birds”)
2. **Pronoun**: a word used in place of a noun (e.g. “it”, “they”, “I”)
3. **Verb**: express action in the sentence (e.g. “sing”)
4. **Adjective**: describe properties of nouns (e.g. “yellow”)
5. **Adverb**: modifies or describes a verb, an adjective, or another adverb (e.g. “extremely”, “slowly”)
6. **Preposition**: a word placed before a noun/pronoun to form a phrase modifying another word/phrase (e.g. “in”, “for”, “without”)

SYNTAX AND PARTS OF SPEECH



sentence level

sentence

phrase level

noun phrase

verb phrase

noun phrase

word level

determiner noun verb determiner noun

The *students* *read* *a* *book*

SEMANTICS

- Semantics is concerned with the meaning of words

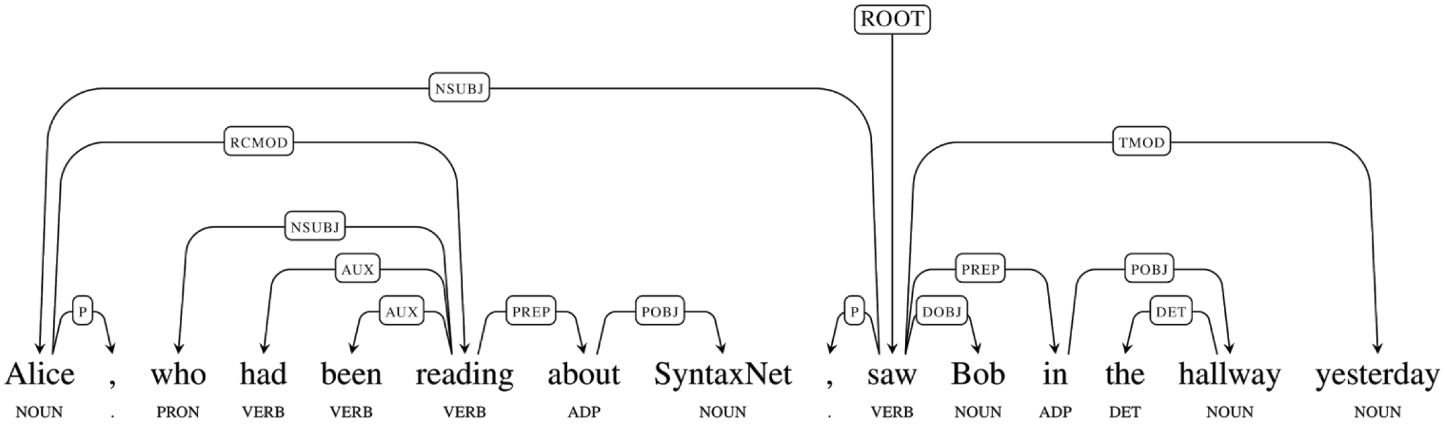
Sense Relation	**Description**	**Example**
Synonymy	words having the same or similar meanings	to buy & to purchase / bill & invoice
Homonymy	a word with multiple unrelated meanings	to order / application
Polysemy	a word with multiple related meanings	to acquire / table
Hyponymy	type-of relationship between words	to build & to create / carrot & vegetable
Meronymy	part-of relationship between words (nouns only)	finger & hand

AGENDA

1. Why Natural Language Processing?
2. Background on Linguistics
3. Technology for Analyzing Natural Language Documents
4. NLP in Process Models
5. NLP in Textual Process Descriptions
6. NLP in Event Logs

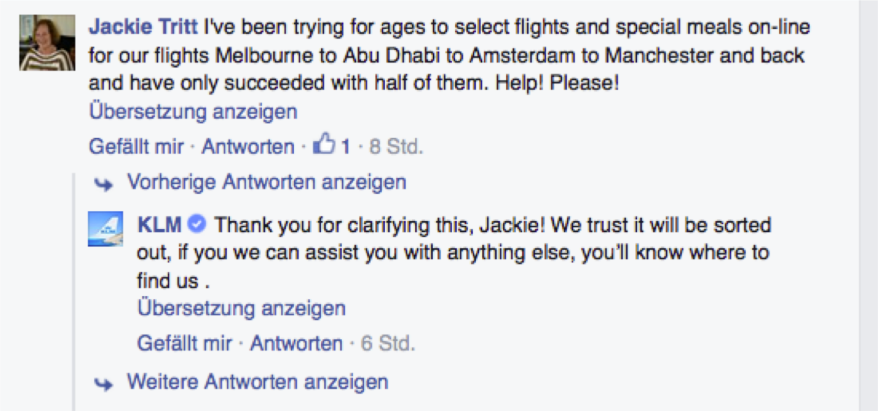
THERE IS POWERFUL NLP TECHNOLOGY AVAILABLE

- https://nlp.stanford.edu/software/lex-parser.html



HOW TO REPRESENT A DOCUMENT?

- Represent by a string?
  - No semantic meaning
- Represent by a list of sentences?
  - Sentence is just like a short document (recursive definition)



Jackie Tritt I've been trying for ages to select flights and special meals on-line for our flights Melbourne to Abu Dhabi to Amsterdam to Manchester and back and have only succeeded with half of them. Help! Please!

Übersetzung anzeigen

Gefällt mir · Antworten · 1 · 8 Std.

Vorherige Antworten anzeigen

KLM Thank you for clarifying this, Jackie! We trust it will be sorted out, if you we can assist you with anything else, you'll know where to find us .

Übersetzung anzeigen

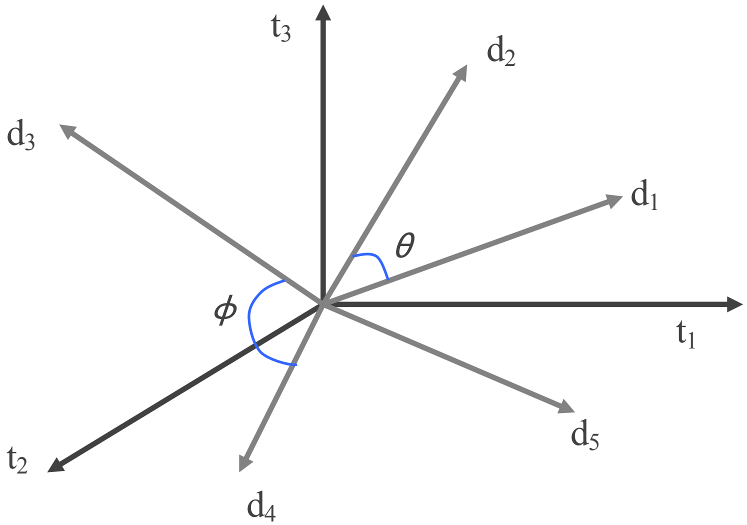
Gefällt mir · Antworten · 6 Std.

Weitere Antworten anzeigen

DOCUMENTS AS VECTORS

- Not all index terms are equally useful in representing document content
- Each document *d* can be viewed as a vector of *weights*
- *The* weights denote the importance of each term
- Terms are axes of vector space
- Documents are points in this vector space

VECTOR SPACE MODEL: INTUITION



Postulate: Documents that are "close together" in the vector space talk about the same things.

BAG-OF-WORDS REPRESENTATION

Term as the basis for vector space

- d_1: Text mining is to identify useful information.
- d_2: Useful information is mined from text.
- d_3: Apple is delicious.

	text	information	identify	mining	mined	is	useful	to	from	apple	delicious
d_1	1	1	1	1	0	1	1	1	0	0	0
d_2	1	1	0	0	1	1	1	0	1	0	0
d_3	0	0	0	0	0	1	0	0	0	1	1

But: Boolean representation is too simple …

HOW TO ASSIGN WEIGHTS?

- Why?
  - Corpus-wise: some terms carry more information about the document content
  - Document-wise: not all terms are equally important

- How?
  - Two basic heuristics
    - TF (Term Frequency) = Within-doc-frequency
    - IDF (Inverse Document Frequency)

TERM FREQUENCY

- A *term* is a word, or some other frequently occurring item
- Given a term *t*, a document *d*, the the term frequency (TF) for *t* is:

$$\text{TF (t,d)} = \frac{\text{Number of times } t \text{ appears in } d}{\text{Total number of terms in the document}}$$

n_{ij}

d_1: Text mining is to identify useful information.

$$\text{TF ("is",}d_1) = \frac{1}{7} = 0.14$$

INVERSE DOCUMENT FREQUENCY

- Term frequency (TF) is a measure of the importance of term *t* in document *d*
- Inverse document frequency (IDF) is a measure of the *general* importance of the term
- For example a high term frequency of “apple” means that “apple” is an important word in a specific document
- A high document frequency (low inverse document frequency) for “apple”, given a particular set of documents, means that “apple” is not all that important overall, since it is in all of the documents

INVERSE DOCUMENT FREQUENCY

- Given a collection of documents *D*, the inverse document frequency (IDF) of a term *t* is:

$$\mathrm{IDF}\ (t,D) = \log \frac{|D|}{|d \in D: t \in d|}$$

- Log of: … the number of documents in the collection, divided by the number of those documents that contain the term

$$\mathrm{TFIDF}\ (t,d,D) = \mathrm{tf}(t,d) * \mathrm{IDF}(t,D)$$

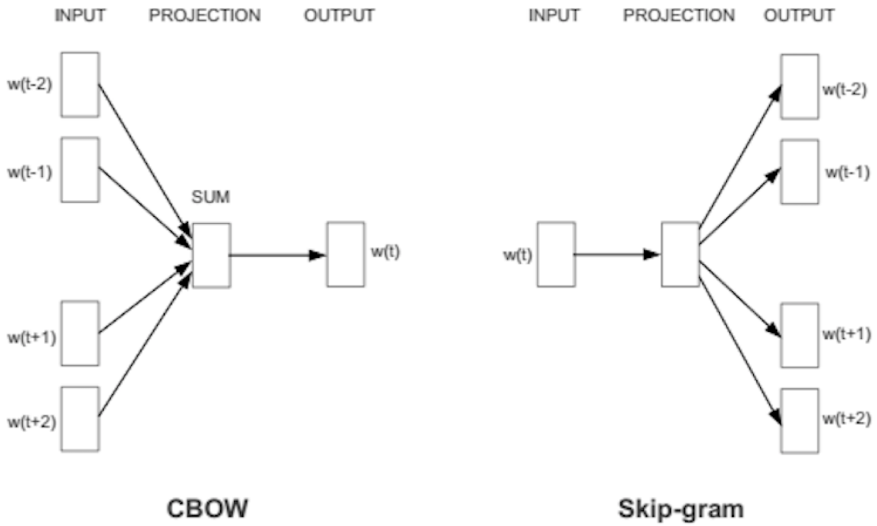
TF/IDF REPRESENTATION

- d_1: Text mining is to identify useful information.
- d_2: Useful information is mined from text.
- d_3: Apple is delicious.

	text	information	identify	mining	mined	is	useful	to	from	apple	delicious
d_1	0.14* 0.18	0.14* 0.18	0.14* 0.48	0.14* 0.48	0	0.14* 0	0.14* 0.18	0.14* 0.48	0	0	0
d_2	0.16* 0.18	0.16* 0.18	0	0	0.16* 0.48	0.16* 0	0.16* 0.18	0	0.16* 0.48	0	0
d_3	0	0	0	0	0	0.33* 0	0	0	0	0.33* 0.48	0.33* 0.48

WORD2VEC

- Predict words using context
- Two versions: CBOW (continuous bag of words) and Skip-gram



https://skymind.ai/wiki/word2vec

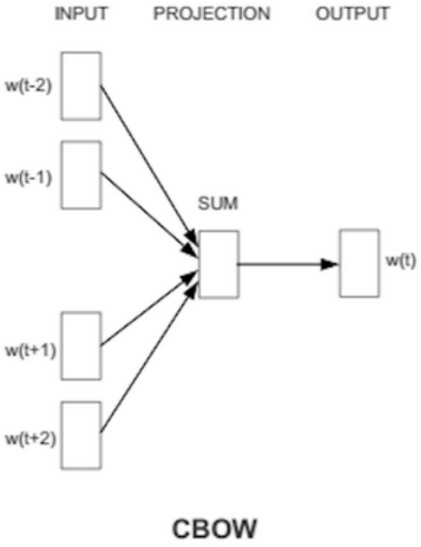
CBOW

Bag of words

- Gets rid of word order. Used in discrete case using counts of words that appear.

CBOW

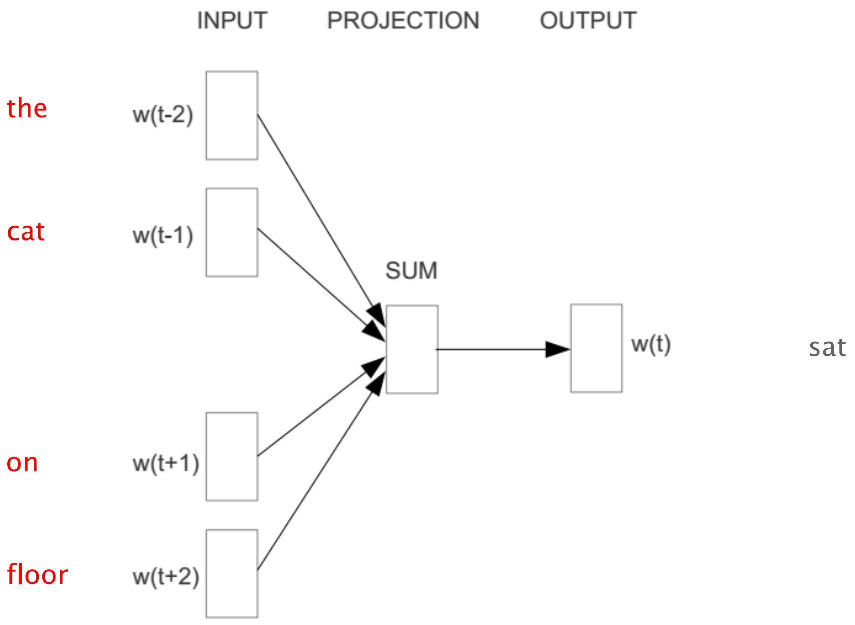
- Takes vector embeddings of n words before target and n words after and adds them (as vectors).
- Also removes word order, but the vector sum is meaningful enough to deduce missing word.



WORD2VEC – CONTINUOUS BAG OF WORD

E.g. “The cat sat on floor”

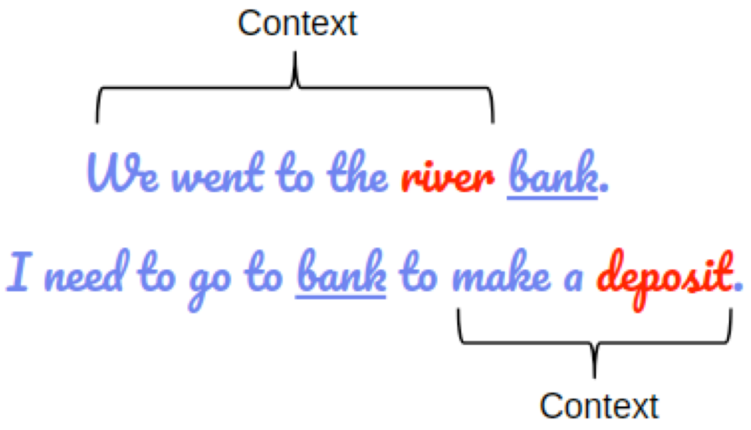
- Window size = 2



www.cs.ucr.edu/~vagelis/classes/CS242/slides/word2vec.pptx

BERT

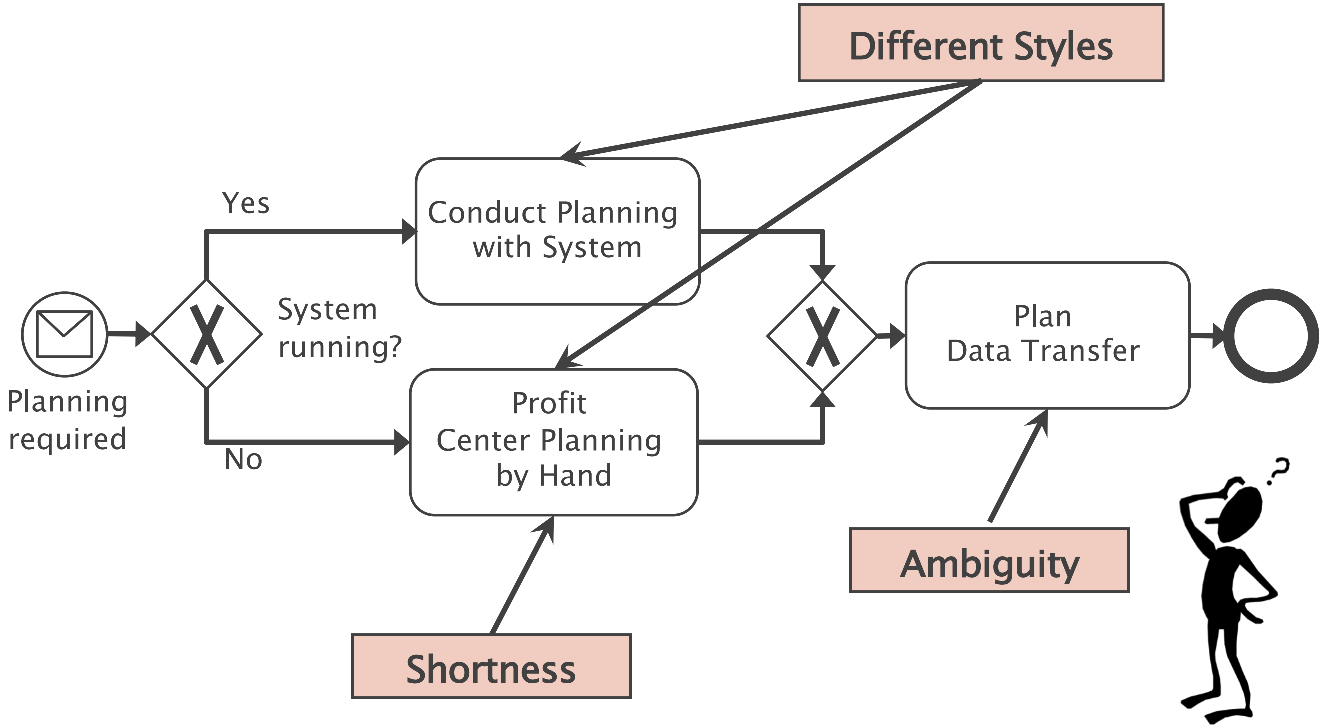
- ”BERT = Bidirectional Encoder Representations from Transformers
- Designed to pre-train deep bidirectional representations from unlabeled text by jointly conditioning on both left and right context
- The pre-trained BERT model can be fine-tuned with just one additional output layer to create state-of-the-art models for a wide range of NLP tasks


Context
We went to the river bank.
I need to go to bank to make a deposit.
Context

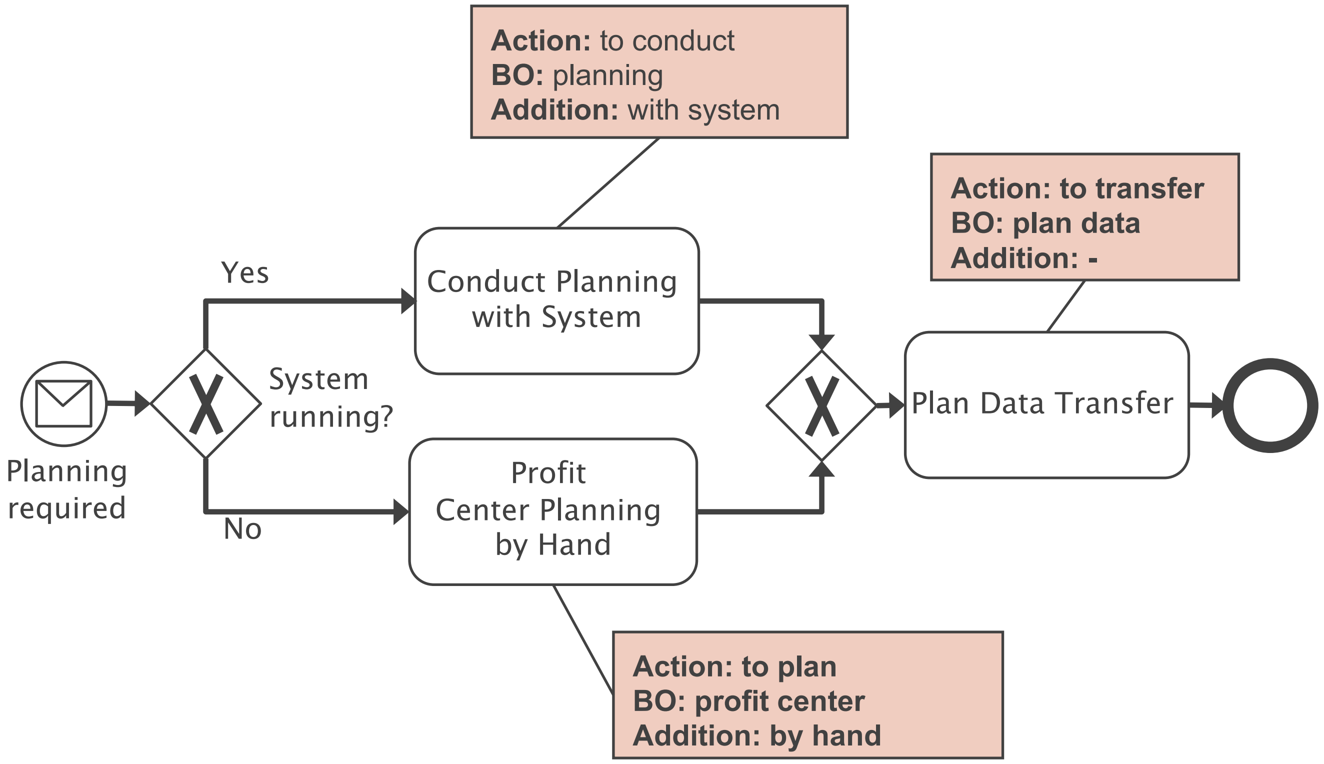
AGENDA

1. Why Natural Language Processing?
2. Background on Linguistics
3. Technology for Analyzing Natural Language Documents
4. NLP in Process Models
5. NLP in Textual Process Descriptions
6. NLP in Event Logs

CHALLENGES OF ANALYZING LANGUAGE IN PROCESS MODELS



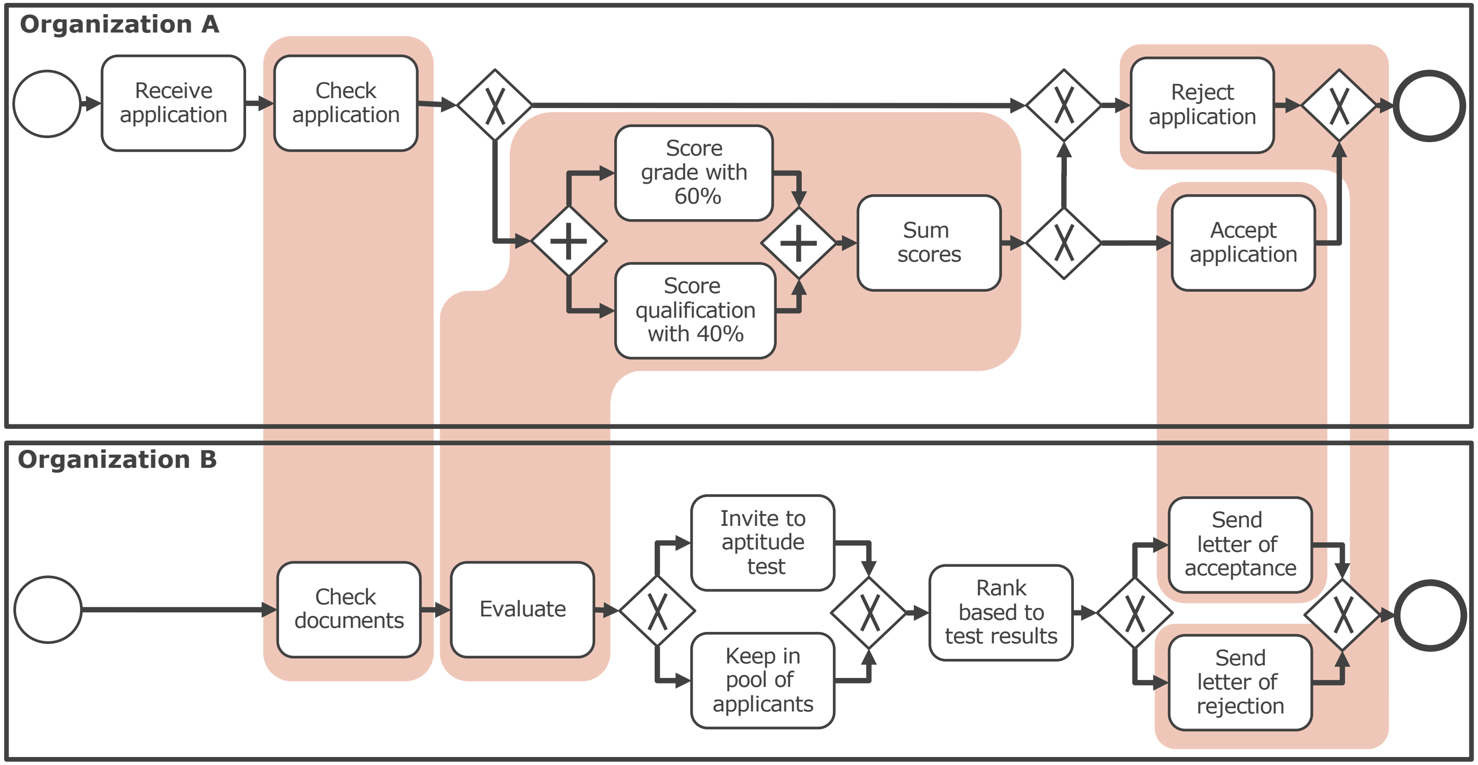
GOAL OF ANALYSIS: ANNOTATION



[1] H. Leopold, S. Smirnov, J. Mendling: **On the Refactoring of Activity Labels in Business Process Models.** Information Systems 37(5): 443–459, 2012.

[2] H. Leopold, H. van der Aa, J. Offenberg, H. A. Reijers: **Using Hidden Markov Models for the Accurate Linguistic Analysis of Process Model Activity Labels.** Information Systems 83: 30–39, 2019.

PROCESS MODEL MATCHING



[3] C. Meilicke, H. Leopold, E. Kuss, H. Stuckenschmidt, H. A. Reijers: **Overcoming Individual Process Model Matcher Weaknesses Using Ensemble Matching.** Decision Support Systems 100: 15-26, 2017.

[4] H. Leopold, M. Niepert, M. Weidlich, J. Mendling, R. M. Dijkman, H. Stuckenschmidt: **Probabilistic Optimization of Semantic Process Model Matching.** In: 10th International Conference on Business Process Management (BPM 2012), September 3-6, 2012, Tallinn, Estonia.

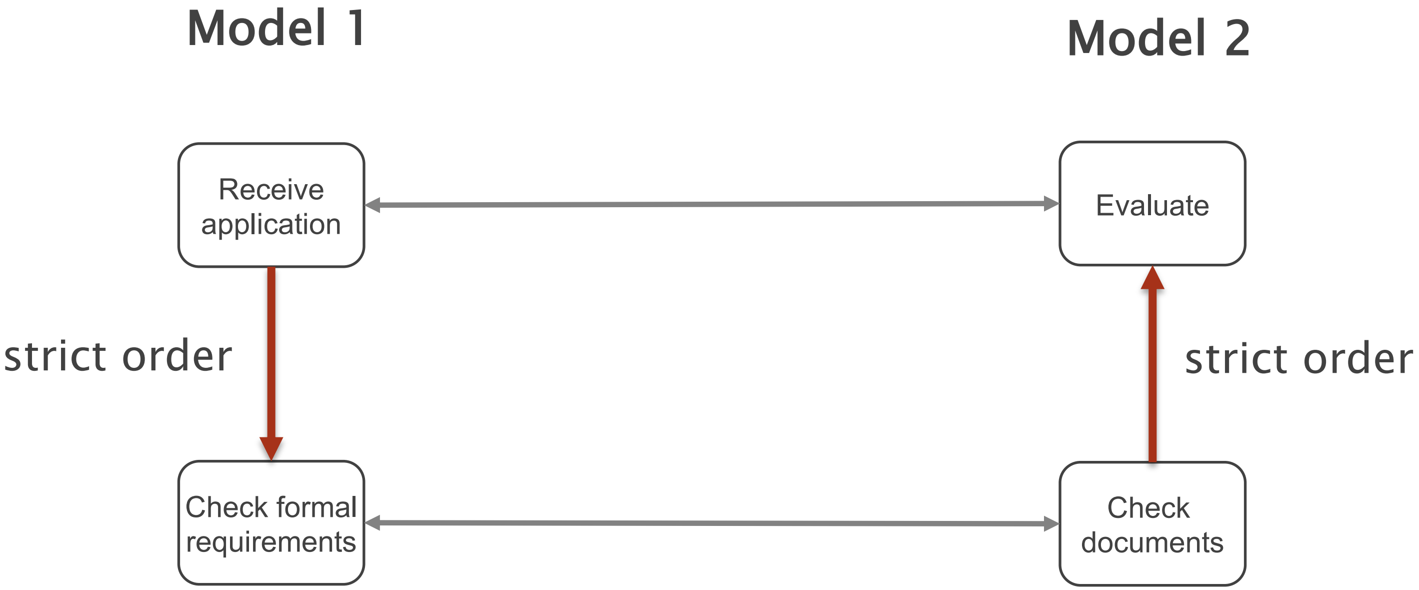
GENERATION OF SEMANTIC MATCH HYPOTHESES

Idea: Calculate semantic similarity of components

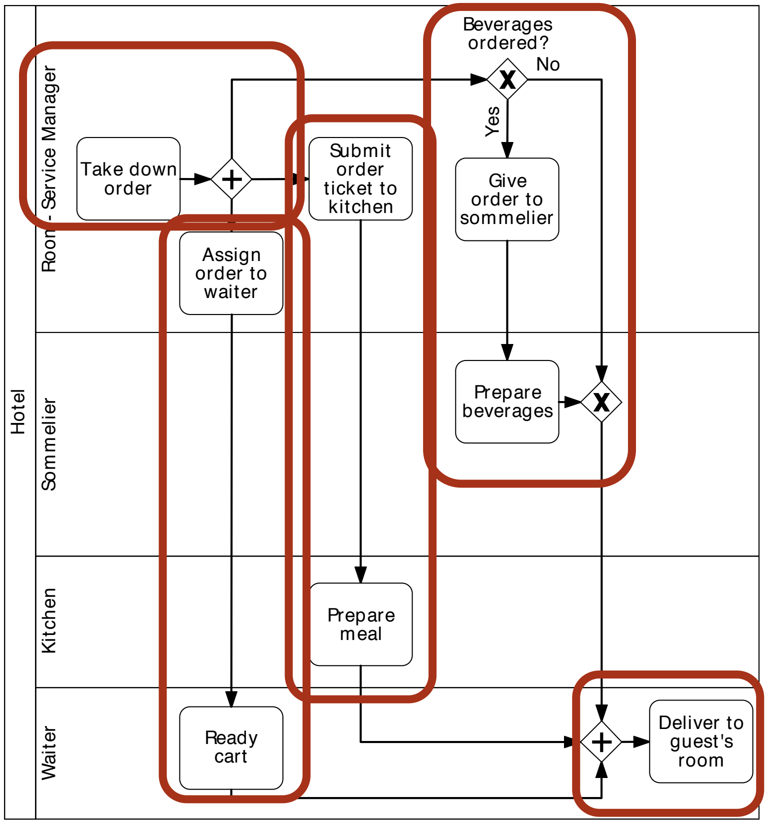
	Assess records	Evaluate documents	**Similarity (Lin)**
action	assess	evaluate	1.0
object	records	documents	0.73

→ Similarity score of activities is **0.87**

PROBABILISTIC MATCH OPTIMIZATION



PROCESS VERBALIZATION



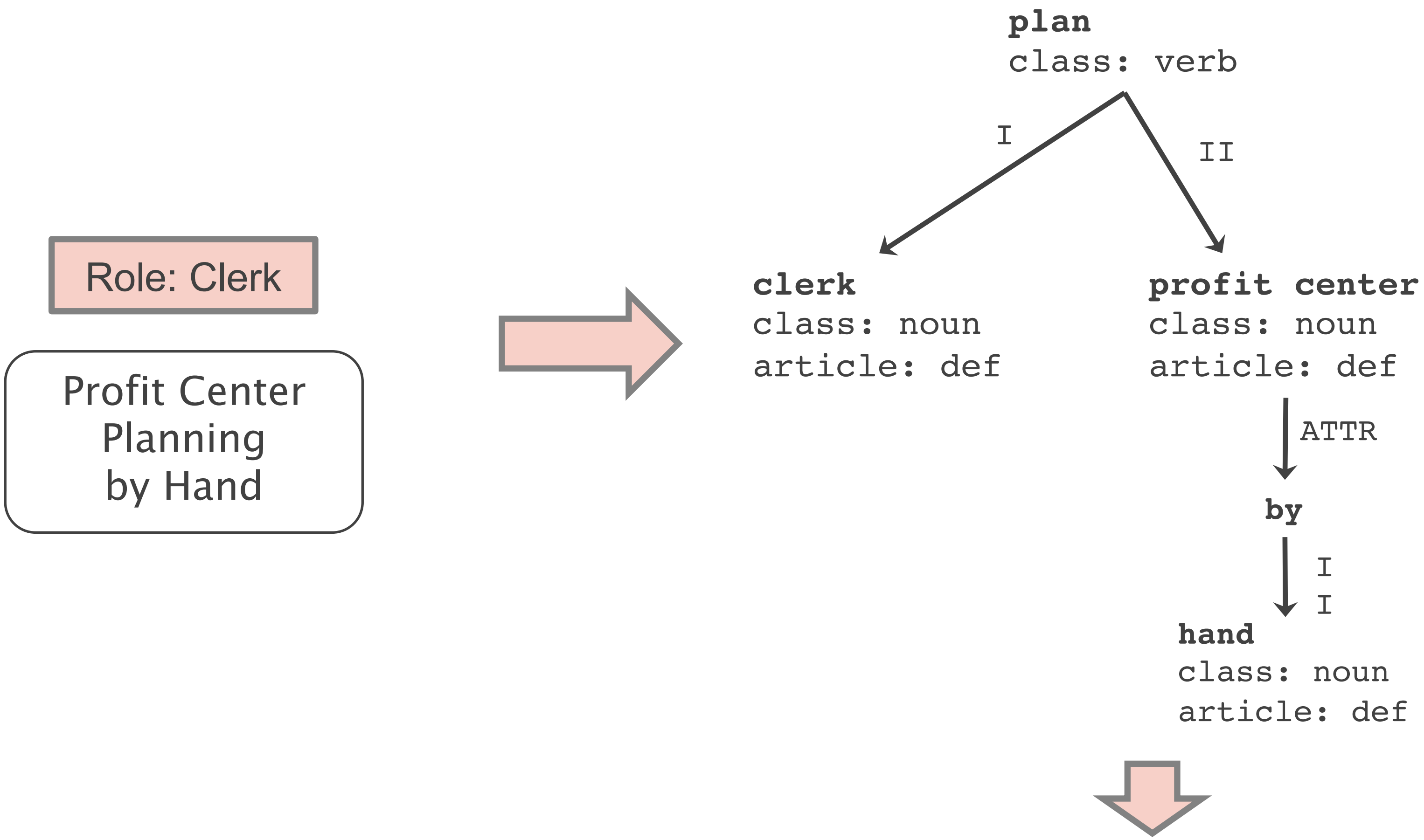
The process begins when the Room-Service Manager takes down an order. Then, the process is split into 3 parallel branches:

- In case beverages are ordered, the Room-Service Manager gives the order to the sommelier. Afterwards, the Sommelier prepares the beverages.
- The Room-Service Manager assigns the order to the waiter. Subsequently, the Waiter readies the cart.
- The Room-Service Manager submits the order ticket to the kitchen. Then,the Kitchen prepares the meal.

Once all 3 branches were executed, the Waiter delivers to the guest's room and the process is finished.

[5] H. Leopold, J. Mendling, A. Polyvyanyy: **Supporting Process Model Validation through Natural Language Generation**. IEEE Transactions on Software Engineering 40(8): 818-840, 2014.

PROCESS VERBALIZATION - DETAILS



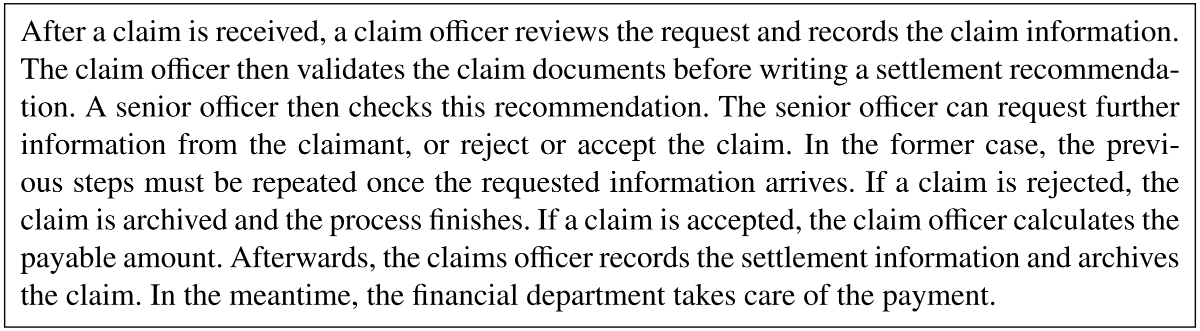
"The clerk plans the profit center by hand."

AGENDA

1. Why Natural Language Processing?
2. Background on Linguistics
3. Technology for Analyzing Natural Language Documents
4. NLP in Process Models
5. NLP in Textual Process Descriptions
6. NLP in Event Logs

TEXTUAL PROCESS DESCRIPTIONS

- Automated business process analysis techniques provide valuable opportunities to organizations
- They cannot be simply applied to textual process descriptions

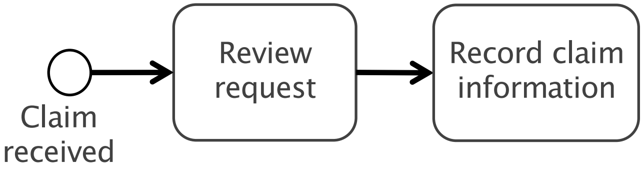


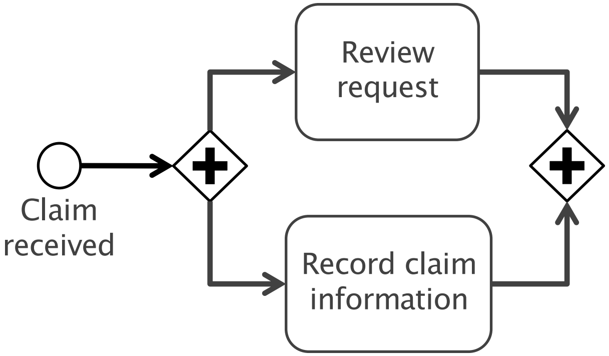
After a claim is received, a claim officer reviews the request and records the claim information. The claim officer then validates the claim documents before writing a settlement recommendation. A senior officer then checks this recommendation. The senior officer can request further information from the claimant, or reject or accept the claim. In the former case, the previous steps must be repeated once the requested information arrives. If a claim is rejected, the claim is archived and the process finishes. If a claim is accepted, the claim officer calculates the payable amount. Afterwards, the claims officer records the settlement information and archives the claim. In the meantime, the financial department takes care of the payment.

BEHAVIORAL AMBIGUITY (1/2)

“*After a claim is received, a claims officer reviews the request* ***and*** *records the claim information.*”

- Q: Is it OK if the claims officer records the claim before reviewing the request?

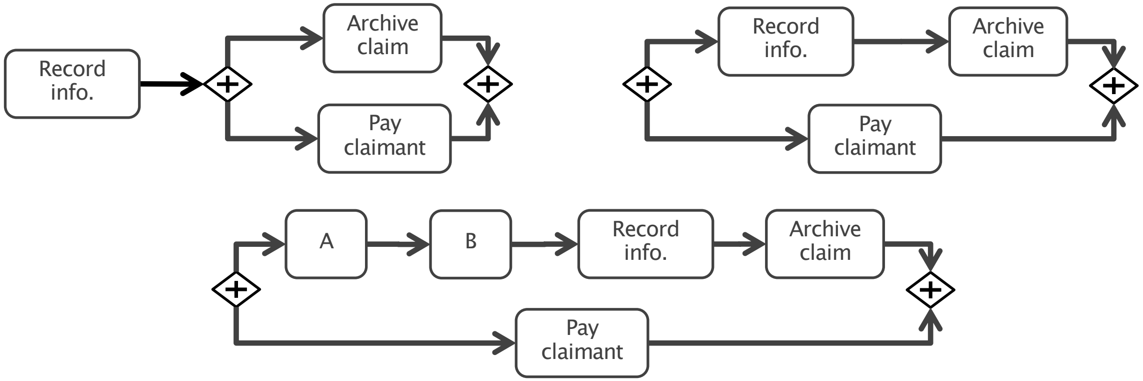




BEHAVIORAL AMBIGUITY (2/2)

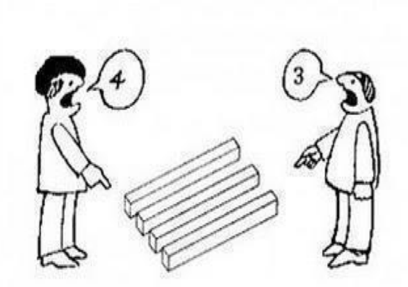
"[…] *Afterwards, the claims officer records the settlement information and archives the claim.* ***In the meantime****, the financial department takes care of the payment.*"

- Q: When can the financial department start to pay the claimant?



CHALLENGE OF AMBIGUITY

- Ambiguity in textual process descriptions can lead to different views on how to properly carry out a business process

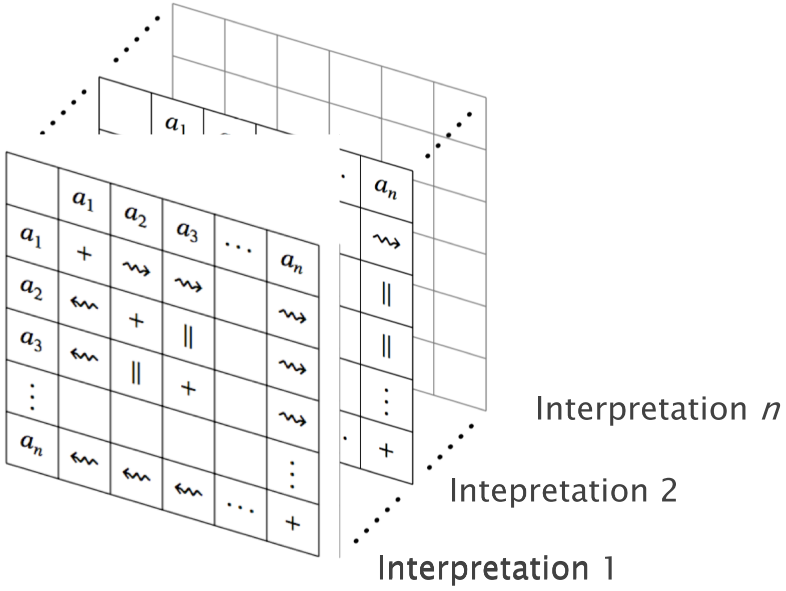


- Presents a challenge to techniques for automated business process analysis

- **Problem:** imposing assumptions on the interpretation of textual descriptions can lead to incorrect conclusions

BEHAVIORAL SPACE

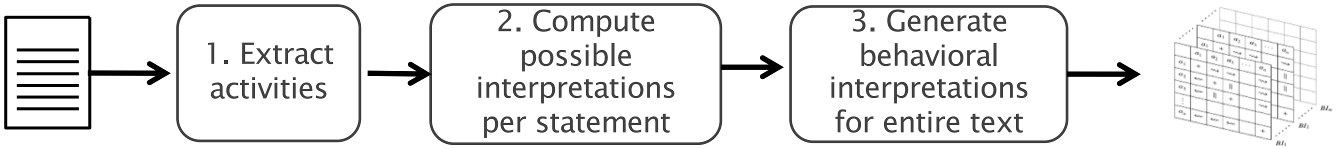
- Different way to dealing with ambiguity
- Capture causes and effects of ambiguity in a structured manner
- Allows for reasoning, without the need to impose assumptions



	a_1	a_2	a_3	$\dots$	a_n
a_1	$+$	$\rightsquigarrow$	$\rightsquigarrow$		$\rightsquigarrow$
a_2	$\leftsquigarrow$	$+$	$\parallel$		$\rightsquigarrow$
a_3	$\leftsquigarrow$	$\parallel$	$+$		$\rightsquigarrow$
$\vdots$					$\vdots$
a_n	$\leftsquigarrow$	$\leftsquigarrow$	$\leftsquigarrow$	$\dots$	$+$

[5] H. van der Aa, H. Leopold, H. A. Reijers: **Checking Process Compliance against Natural Language Specifications using Behavioral Spaces.** Information Systems 78: 83-95, 2018.

AUTOMATICALLY OBTAINING BEHAVIORAL SPACES


1. Extract activities

2. Compute possible interpretations per statement

3. Generate behavioral interpretations for entire text

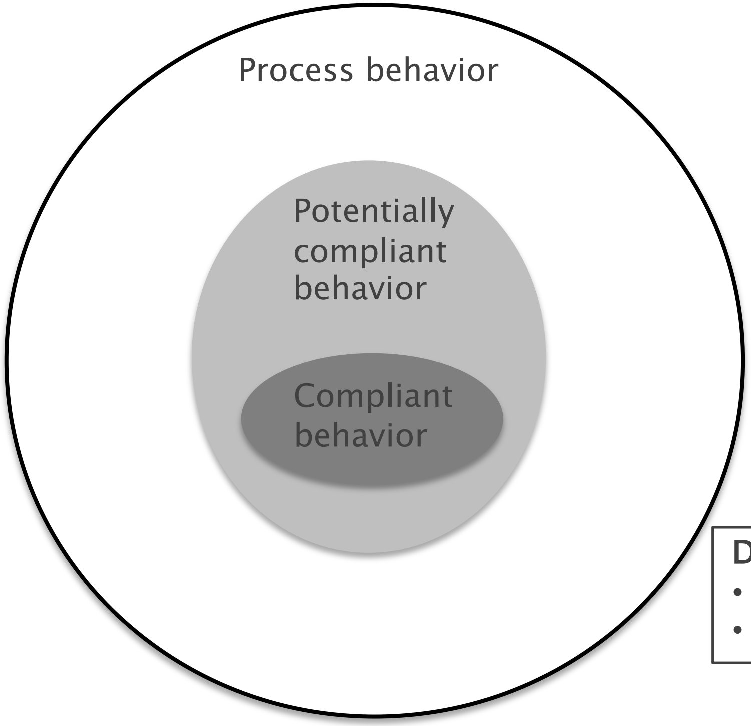
> S_1: "***Afterwards***, *the claims officer records the settlement information* ***and*** *archives the claim.*"

- a_1 = record settlement info., a_2 = archive claim
- $S_1^1 = \{a_0 \rightarrow a_1, a_0 \rightarrow a_2, a_1 \rightarrow a_2\}$
- $S_1^2 = \{a_0 \rightarrow a_1, a_0 \rightarrow a_2, a_1 \parallel a_2\}$

> S_2: "***In the meantime***, *the financial department takes care of the payment.*"

- a_3 = pay claimant
- $S_2^1 = \{a_0 \rightarrow a_3, a_1 \rightarrow a_3, a_2 \parallel a_3\}$
- $S_2^2 = \{a_0 \rightarrow a_3, a_1 \parallel a_3, a_2 \parallel a_3\}$
- $S_2^3 = \{a_0 \parallel a_3, a_1 \parallel a_3, a_2 \parallel a_3\}$

BEHAVIORAL SPACE COMPLIANCE



a_1: record settlement info.
a_2: archive claim
a_3: pay claimant

t_1: $<a_1, a_2, a_3>$ ✓

t_2: $<a_2, a_1, a_3>$ ✗

t_3: $<a_3, a_1, a_2>$?

Diagnostics:

- Complies to 2/3 of interpretations;
- Only if statement s_2 implies $a_1||a_3$

AGENDA

1. Why Natural Language Processing?
2. Background on Linguistics
3. Technology for Analyzing Natural Language Documents
4. NLP in Process Models
5. NLP in Textual Process Descriptions
6. NLP in Event Logs

PROBLEM AND GOAL

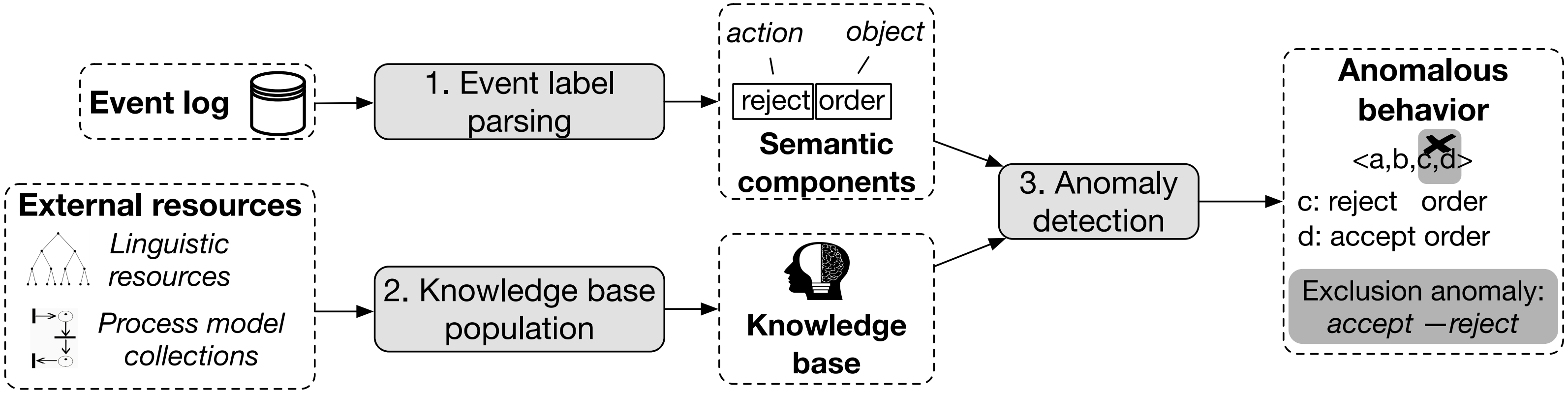
Detecting anomalies in event logs:

- Out-of-order execution
- Superfluous event
- Missing event

Trace 1	
A	Create order
C	Approve order
B	Check order
E	Create delivery
F	Complete delivery

Trace 2	
A	Create order
B	Check order
C	Approve order
D	Reject order
F	Complete delivery

PROPOSED SOLUTION



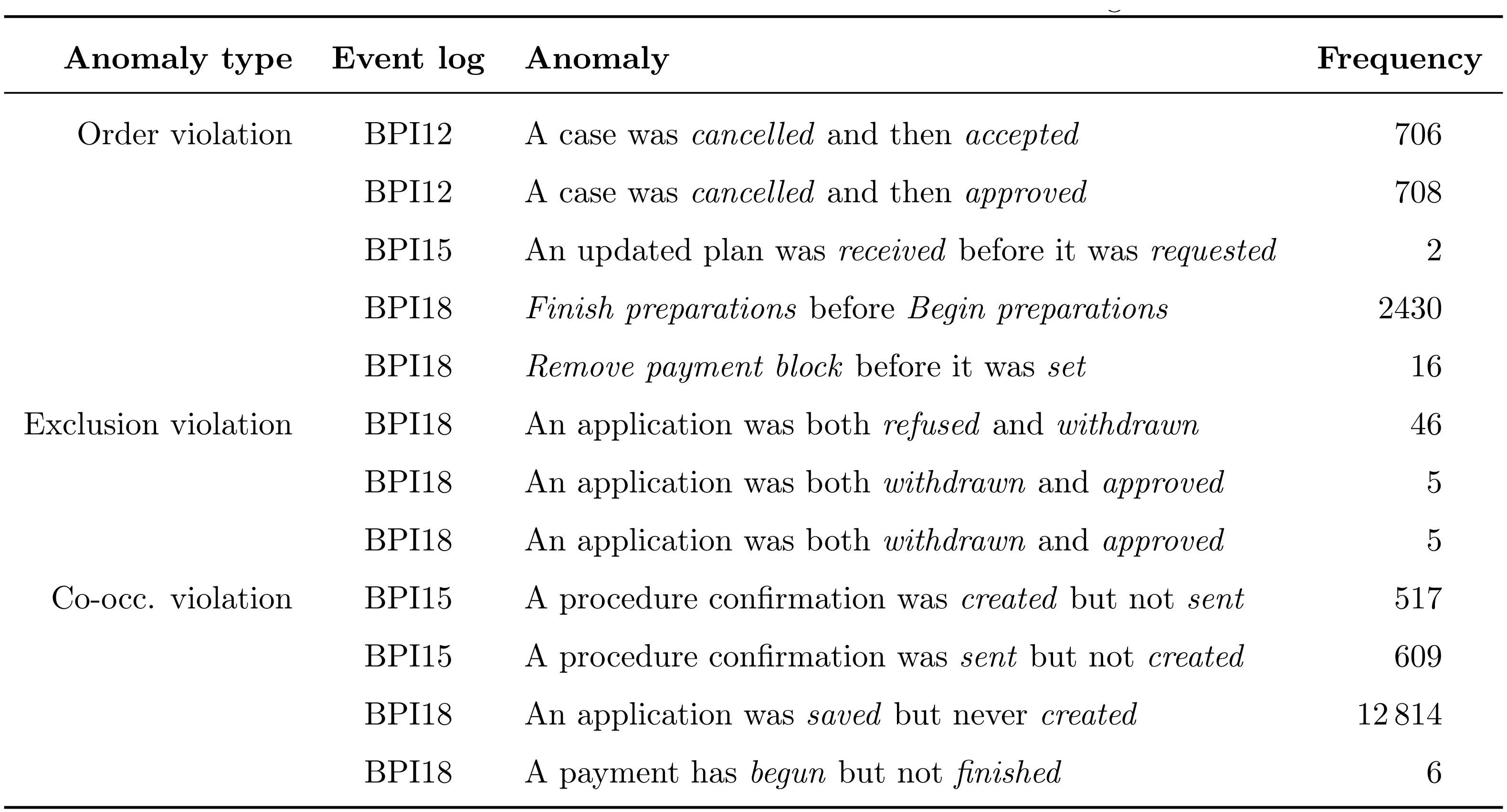
1. **Event label parsing**: Extract semantic information from event log labels (action + business object)
2. **Knowledge base population**: Collect assertions about the interrelations that should hold among actions applied to business objects
3. **Anamoly detection**: Compare the semantic information from label parsing to the assertions contained in the knowledge base in order to detect semantically anomalous behavior

[6] H. van der Aa, A. Rebmann, H. Leopold: **Natural Language-based Detection of Semantic Execution Anomalies in Event Logs.** Information Systems, 2021.

KNOWLEDGE BASE POPULATION

- There are many resources that capture semantic relations between words: WordNet, VerbNet, DBpedia, VerbOcean
- For our purposes, VerbOcean is most promising:
  - **Happens-before**:
    - (consider, $\prec$, decide)
    - (bill, $\prec$, reimburse)
  - **Enablement**:
    - (make, $\Rightarrow$, sell)
    - (compare, $\Rightarrow$, describe)
  - **Antonymy**:
    - (accept, #, reject)
    - (acquire, #, lose)
- Of course, such relations can also be mined from process repositories ...

SOME RESULTS



Anomaly type	Event log	Anomaly	Frequency
Order violation	BPI12	A case was *cancelled* and then *accepted*	706
	BPI12	A case was *cancelled* and then *approved*	708
	BPI15	An updated plan was *received* before it was *requested*	2
	BPI18	*Finish preparations* before *Begin preparations*	2430
	BPI18	*Remove payment block* before it was *set*	16
Exclusion violation	BPI18	An application was both *refused* and *withdrawn*	46
	BPI18	An application was both *withdrawn* and *approved*	5
	BPI18	An application was both *withdrawn* and *approved*	5
Co-occ. violation	BPI15	A procedure confirmation was *created* but not *sent*	517
	BPI15	A procedure confirmation was *sent* but not *created*	609
	BPI18	An application was *saved* but never *created*	12 814
	BPI18	A payment has *begun* but not *finished*	6

SUMMARY

- NLP provides us with many interesting opportunities for analyzing process models, textual process descriptions, and event logs
- There are many unsolved challenges where NLP can help:
  - Event log extraction from databases
  - Analzying user interactions logs
  - Inferring process improvement strategies from social media posts
  - ...
- NLP techniques are never perfectly accurate but in many cases they enable analyses that otherwise would be simply impossible