

26 January 2018 De Vereeniging, Nijmegen

CLIN 28

 ${\bf Book\ of\ abstracts\ for\ the}$ 28th meeting of Computational Linguistics in the Netherlands

Nijmegen, 26 January 2018

Automagically generated and typeset in LATEX.
Book of abstracts of the 28th Meeting of Computational Linguistics in the Netherlands (CLIN 28).
De Vereeniging, Nijmegen
⊕2018 https://github.com/LanguageMachines/clin28 Book: louis
Font logo: Final Frontier Old Style by Allen R. Walden Contact: clin28@cls.ru.nl

Preface

Welcome to the 28th Computational Linguistics in The Netherlands meeting!

Yet again, CLIN, the non-centrally organised yearly meeting of computational linguists from the Low Countries, returns to Nijmegen, one of the pillars of computational linguistics in the Netherlands since the early 1970s. In large brushstrokes, Nijmegen was the home of Jan Aarts' corpus linguistics group TOSCA, Jan van Bakel's more formally oriented computational linguistics group, Gerard Kempen's all-round cognitive and computational psycholinguistics group, Theo van den Heuvel's Polderland, and Lou Boves' speech technology group. Presently, work in computational linguistics and speech technology happens at the Centre for Language and Speech Technology (formerly SPEX) within the Faculty of Arts of Radboud University. This work echoes that of their predecessors: it is closely tied with speech technology research, and while being characterised mostly as data-driven and empirical, the research maintains strong ties to theoretical linguistics. Computational linguistics in Nijmegen continues to profit, furthermore, of the presence of psycholinguistic and sociolinguistic expertise, a strong communication and information science group, and the on-campus research in artificial intelligence, data science, and more recently neuroscience.

Welcome to Nijmegen!

Concertgebouw De Vereeniging, the centrally located 1916 concert hall inviting architecture buffs to spot its Art Deco and Art Nouveau elements, offers the surroundings of this year's CLIN. We are happy and proud to present prof. Andy Way (Dublin City University) as keynote speaker who will talk about machine translation. It will be an interesting challenge for you, as usual, to find an optimal way through the interesting program with 44 oral presentations in four parallel sessions (one of which will be the winner of the STIL Master Thesis Prize) and 40 poster presentations.

Many thanks are due to the Centre for Language and Speech Technology and to our sponsors for their contribution to CLIN28, and to the organising team: Florian Kunneman, Iris Hendrickx, Merijn Beeksma, Frank Grootjen, Erkan Basar, Alessandro Lopopolo, Chara Tsoukala, Christoph Aurnhammer, Nelleke Oostdijk, Eric Sanders, Louis Onrust, Maarten van Gompel, and Hans van Halteren.

In the wake of CLIN28 a call for submissions will be sent to all contributors of accepted CLIN28 contributions for the 8th Volume of the CLIN Journal, which will publish a peer-reviewed selection of papers of this year's CLIN.

Enjoy CLIN28!

Antal van den Bosch

Committees

Organising Committee

Christoph Aurnhammer, M. Erkan Basar, Merijn Beeksma, Antal van den Bosch, Franc Grootjen, Iris Hendrickx, Florian Kunneman, Alessandro Lopopolo, louis onrust and Chara Tsoukala.

Support

Maarten van Gompel, Hans van Halteren, Ian Hancock, Ali Hürriyetoğlu, Yung Han Khoe, Ekaterina Slesareva, Eric Sanders, Odette Scharenborg, Ko van der Sloot, Tim Voets and Peter van de Water.

Overview of CLIN meetings

1990	October 26	OTS Utrecht
1991	November 29	CWI Amsterdam
1992	October 30	ITK Tilburg
1993	November 25	Alfa-informatica Groningen
1994	November 23	TU Twente
1995	December 1	Antwerpen
1996	November 15	IPO Eindhoven
1997	December 12	Letteren KUN, Nijmegen
1998	December 11	CCL, KU Leuven
1999	December 10	OTS Utrecht
2000	November 3	Taal en Informatica, Tilburg
2001	November 30	Computer Science, University of Twente
2002	November 29	Alfa-informatica, University of Groningen
2003	December 19	Centre for Dutch Language and Speech, University of Antwerp
2004	December 17	Leiden Centre for Linguistics
2005	December 16	University of Amsterdam
2007	January 12	CCL, KU Leuven
2007	December 7	Radboud University Nijmegen
2009	January 22	Groningen University
2010	February 5	UiL-OTS, Utrecht University
2011	February 11	LT3, University College Ghent
2012	January 20	ILK, Tilburg University
2013	January 18	HMI, University of Twente, Enschede
2014	January 17	INL & TST-Centrale, Leiden
2015	February 5-6	CLiPS, University of Antwerp
2016	December 18	CLTL, VU Amsterdam
2017	February 10	CCL, KU Leuven
2018	January 26	CLS, Radboud University

Wireless Network

SSID vereeniging_gast Password vereeniging

Gold

/instituut voor de Nederlandse taal/

Silver



Beter zoeken. Zeker weten.

textkernel

Machine Intelligence for People and Jobs

Bronze















Coffee
and
Registration
9:30
3:30 -

	Shared Task and Discussion on Open data Annazaal	Shared task overview	Shared task presentations	Open Data, Social Media and other User Generated Text: some topics for a discussion session Ineke Schuurman, Leen Sevens, Vincent Vandeghinste
	IR and Knowledge Extraction Sociëteitskamer	Lawyers that Search? Thesaurus driven legal search using custom analysis and query parsing and beyond Tjerk de Greef and Pieter van Boxtel	User profiling with personal graphs Suzan Verberne	Methods For Automatically Generating a Legal Thesaurus and discussing NLP in the Legal Domain Hugo de Vos
	Semantics Keizer Karel Foyer	Towards Wide- coverage Composi- tional Semantics for Dutch Johan Bos, Lasha Abzian- idze, Hessel Haagsma and Rik van Noord	From Topic Modeling to Concept Identification. The Added Value of Word Embeddings to Discern Semantic Coherence in Keyword Clusters Mathias Coeckelbergs	IDEA: Idiom Detection – an Easy Approach Anna Feldman, Jing Peng and Katsiaryna Aharodnik
	Language generation Leeuwzaal	Automating lexical simplification in Dutch Bram Bulte, Leen Sevens and Vincent Vandeghinste	Pictograph-to-Text Translation for Augmented and Alternative Communication Leen Sevens, Vincent Vandeghinste, Lyan Verwimp, Ineke Schuurman, Patrick Wambacq and Frank Van Eynde	Content Generation in German and Dutch from Structured Data Andreas Madsack and Jo- hanna Heininger
Centrale Hal	Emotion and Senti- ment	Semi-supervised emotion lexicon expansion with label propagation Mario Giulianelli and Daniël de Kok	Using common sense to detect irony on Twitter Cynthia Van Hee, Els Lefever and Véronique Hoste	OMGhetwaszooolekker!Canfine-grainedsentimentanalysisbenefitfromcoreferenceresolution?tion?OrphéeDeClercqVéroniqueHoste
0.6.9	9:30 - 10:50	9:30	9:50	10:10

Classifying Definitional Sentences using Multi-view Neural Network Architectures Subhradeep Kayal, Sameer Chivukula and Sophia Katrenko							NLP for historical text Annazaal Combining grammatical features for dating texts in small-corpus languages Camil Staps
Lexical Knowledge Acquisition with Theorem Proving Lasha Abzianidze							Parsing and Grammar Sociëteitskamer Lexical Normalization for Neural Network Parsing Rob van der Goot and Gertjan Van Noord
Generating Topically Coherent Text with Recurrent Neural Networks Tim Van de Cruys				aker)			Language Variation Leeuwzaal Fine-grained genre/register recognition: romance fiction in the BNC Hans Van Halteren and Nelleke Oostdijk
Monitoring stance towards vaccination in Twitter messages Florian Kunneman, Liesbeth Mollema, Antal van den Bosch, Mattijs Lambooij, Albert Wong and Hester de Melker	Coffee Break Centrale hal	Welcome Kleine zaal	STIL Thesis Prize Kleine zaal	Andy Way (Invited speaker) Kleine zaal	Lunch Keizer Karel Foyer	Poster Session Centrale Hal	Deep Learning Kleine zaal Unsupervised dense patient representations with interpretable classification decisions Madhumita Sushil, Simon Šuster, Kim Luyckx and Walter Daelemans
10:30	10:50 - 11:10	11:15 - 11:25	11:25 - 11:35	11:35 - 12:40	12:40 - 13:30 Lunch Keizer]	13:30 - 14:30	14:30 - 15:30

Nederlab: Progress and challenges in linguistic enrichment of historical Dutch Texts Katrien Depuydt, Maarten Van Gompel, Hennie Brugman, Jesse de Does and Gosse Bonma	Good lynx, bad lynx: document enrichment for historical ecologists Marieke van Erp, Jesse De Does, Thomas van Goethem and Katrien Depuydt		Translation Annazaal Post-editing a Novel with Neural Machine Translation Antonio Toral
Memory Based Modelling of Dutch Past Tense Regularity Christoph Aurnhammer	TM-MT: the SCATE sliders Tom Vanallemeersch		Named Entity Recognition and Term Extraction Sociëteitskamer Improving named entity resolution for English to Chinese neural machine translation Camiel Colruyt
BasiScript: a 9 million word corpus of texts written in primary school Agnes Tellings, Nelleke Oostdijk, Iris Monster, Franc Grootjen and Antal van den Bosch	Reducing perplexities and word error rates by improving the interpolation factors of a Bayesian skipgram language model Louis Onrust, Hugo Van hamme and Antal van den Bosch		Authorship and Discourse Leeuwzaal Patterns of rhetorical structure in student's writing in different text genres Rianne Conjin, Simon Knight, Roberto Martinez and Menno van Zaanen
SuperNMT: Integrating Supersense and Supertag Features into Neural Machine Translation Eva Vanmassenhove and Andy Way	Towards multilingual sequence labelling using cross-lingual word embeddings Chao Li, Carsten van Weelden, Luigi Lorato, Carsten L. Hansen, Mihai Rotaru and Jakub Zavrel	15:30 - 15:55 Coffee and beverages Keizer Karel Foyer	Social media and winner STIL Thesis Prize Kleine zaal TBA Winner STIL Thesis Prize
14:50	15:10	15:30 - 15:55	15:55 - 17:15 15:55

145 ways to insert punctuation for speech translation Vincent Vandeghinste, Lyan Verwimp, Joris Pelemans and Patrick	Elastic-substitution decoding for Hierar- chical SMT: efficiency, richer search and double labels Gideon Maillette de Buy Wenniger, Khalil Sima'An and Andy Way	To Wikify or to Babelfy for Implicit Translation Evaluation? Iris Hendrickx, Maarten van Gompel and Antal van den Bosch
Geo-temporal named entity recognition in multilingual picture collections Hans Paulussen and Dirk De Hertog	Global Space: A Reconstructive Model of Word Reading Stéphan Tulkens, Dominiek Sandra and Walter Daelemans	Sequential Labelling with Active Learning to Extract Information about Disasters M. Erkan Basar
What can we learn from behavioral biometric data? Barbara Plank	Microportrait detection for identifying stereotypes in Dutch News Antske Fokkens, Ruigrok Nel, Sarah Gagenstein, Wouter Van Atteveldt and Beukeboom Camiel	Discourse lexicons for multiple languages and their use for gender profiling Ben Verhoeven and Walter Daelemans
	Transfer Learning on Cyberbullying Classifi- cation Long Tran	Relevant Information Detection on Twitter Ali Hürriyetoğlu, Nelleke Oostdijk, Antal van den Bosch and Mustafa Erkan Basar
16:15	16:35	16:55

17:15 - 18:15 **Drinks** Keizer Karel Foyer

Keynote

Neural MT: Separating hype from the reality

Machine Translation (MT) has been used by many people for some time now as a productivity tool, with demonstrable success. Recently, a new paradigm – Neural MT (NMT) – has emerged as a contender to replace statistical MT as the new state-of-the-art in the field. While there is no doubt that NMT has enormous potential, we argue that this has been overhyped, a situation which if left unchecked will lead to unrealistic expectations of its capabilities and ultimately to a host of disappointed users. We will demonstrate that this has been seen before with the advent of new approaches to MT, and each time this happens it has the potential to undermine the relationship MT developers need to have with translators, the principal users of the technology. Despite overblown claims regarding its prospects, we argue – as we have done for many years now – that the translation community has little to fear from MT, and that where human input is required, the translator remains the most important link in the chain. Indeed, MT developers rely on translators in a number of ways, and if MT is to improve still further, bilateral partnerships need to be formed between both communities for such advances to be made as fast as possible.

Oral presentations

145 ways to insert punctuation for speech translation

Vincent Vandeghinste, Lyan Verwimp, Joris Pelemans and Patrick Wambacq

Automating lexical simplification in Dutch

Bram Bulte, Leen Sevens and Vincent Vandeghinste

BasiScript: a 9 million word corpus of texts written in primary school

Agnes Tellings, Nelleke Oostdijk, Iris Monster, Franc Grootjen and Antal van den Bosch

 ${\it Classifying Definitional Sentences using Multi-view Neural Network Architectures}$

Subhradeep Kayal, Sameer Chivukula and Sophia Katrenko

Combining grammatical features for dating texts in small-corpus languages

Camil Staps

Comparing and Integrating Machine Learning and Rule Based Microtext Classification Ali Hürriyetoğlu, Nelleke Oostdijk and Antal van den Bosch

Content Generation in German and Dutch from Structured Data

Andreas Madsack and Johanna Heininger

Discourse lexicons for multiple languages and their use for gender profiling

Ben Verhoeven and Walter Daelemans

Elastic-substitution decoding for Hierarchical SMT: efficiency, richer search and double labels
Gideon Maillette de Buy Wenniger, Khalil Sima'An and Andy Way

Fine-grained genre/register recognition: romance fiction in the BNC

Hans Van Halteren and Nelleke Oostdijk

From Topic Modeling to Concept Identification. The Added Value of Word Embeddings to Discern Semantic Coherence in Keyword Clusters

Mathias Coeckelbergs

Generating Topically Coherent Text with Recurrent Neural Networks

Tim Van de Cruys

Geo-temporal named entity recognition in multilingual picture collections

Hans Paulussen and Dirk De Hertog

Global Space: A Reconstructive Model of Word Reading

Stéphan Tulkens, Dominiek Sandra and Walter Daelemans

Good lynx, bad lynx: document enrichment for historical ecologists

Marieke van Erp, Jesse De Does, Thomas van Goethem and Katrien Depuydt

IDEA: Idiom Detection - an Easy Approach

Anna Feldman, Jing Peng and Katsiaryna Aharodnik

Improving named entity resolution for English to Chinese neural machine translation

Camiel Colruyt

Lawyers that Search? Thesaurus driven legal search using custom analysis and query parsing and beyond

Tjerk de Greef and Pieter van Boxtel

Lexical Knowledge Acquisition with Theorem Proving

Lasha Abzianidze

Lexical Normalization for Neural Network Parsing

Rob van der Goot and Gertjan Van Noord

Memory Based Modelling of Dutch Past Tense Regularity

Christoph Aurnhammer

Methods For Automatically Generating a Legal Thesaurus and discussing NLP in the Legal Domain Hugo de Vos

Microportrait detection for identifying stereotypes in Dutch News

Antske Fokkens, Ruigrok Nel, Sarah Gagenstein, Wouter Van Atteveldt and Beukeboom Camiel

Monitoring stance towards vaccination in Twitter messages

Florian Kunneman, Liesbeth Mollema, Antal van den Bosch, Mattijs Lambooij, Albert Wong and Hester de Melker

Nederlab: Progress and challenges in linguistic enrichment of historical Dutch Texts

Katrien Depuydt, Maarten Van Gompel, Hennie Brugman, Jesse de Does and Gosse Bouma

OMG het was zooo lekker! Can fine-grained sentiment analysis benefit from coreference resolution?

Orphée De Clercq and Véronique Hoste

Open Data, Social Media and other User Generated Text: some topics for a discussion session Ineke Schuurman, Leen Sevens, Vincent Vandeghinste

Patterns of rhetorical structure in student's writing in different text genres

Rianne Conijn, Simon Knight, Roberto Martinez and Menno van Zaanen

 $Pictograph-to\text{-}Text\ Translation\ for\ Augmented\ and\ Alternative\ Communication$

Leen Sevens, Vincent Vandeghinste, Lyan Verwimp, Ineke Schuurman, Patrick Wambacq and Frank Van Eynde

Post-editing a Novel with Neural Machine Translation

Antonio Toral

Reducing perplexities and word error rates by improving the interpolation factors of a Bayesian skipgram language model

Louis Onrust, Hugo Van Hamme and Antal van den Bosch

Relevant Information Detection on Twitter

Ali Hürriyetoğlu, Nelleke Oostdijk, Antal van den Bosch and Mustafa Erkan Basar

Semi-supervised emotion lexicon expansion with label propagation

Mario Giulianelli and Daniël de Kok

Sequential Labelling with Active Learning to Extract Information about Disasters

M. Erkan Basar

SuperNMT: Integrating Supersense and Supertag Features into Neural Machine Translation Eva Vanmassenhove and Andy Way

 $TM ext{-}MT$: the SCATE sliders

Tom Vanallemeersch

To Wikify or to Babelfy for Implicit Translation Evaluation?

Iris Hendrickx, Maarten van Gompel and Antal van den Bosch

Towards Wide-coverage Compositional Semantics for Dutch

Johan Bos, Lasha Abzianidze, Hessel Haagsma and Rik van Noord

Towards multilingual sequence labelling using cross-lingual word embeddings

Chao Li, Carsten van Weelden, Luigi Lorato, Carsten L. Hansen, Mihai Rotaru and Jakub Zavrel

Transfer Learning on Cyberbullying Classification

Long Tran

Unsupervised dense patient representations with interpretable classification decisions Madhumita Sushil, Simon Šuster, Kim Luyckx and Walter Daelemans

User profiling with personal graphs

Suzan Verberne

Using common sense to detect irony on Twitter

Cynthia Van Hee, Els Lefever and Véronique Hoste

What can we learn from behavioral biometric data?

Barbara Plank

Posters

A Lexicon of Dutch Connectives

Peter Bourgonje, Jacqueline Evers-Vermeul, Jet Hoek, Gisela Redeker, Ted Sanders and Manfred Stede

A connectionist model of serial order applied to speaking rate control

Joe Rodd, Hans Rutger Bosker, Mirjam Ernestus and Louis Ten Bosch

A filter for syntactically comparable parallel sentences

Martin Kroon and Sjef Barbiers

An NLP workflow for replicability

Minh Le

A practical linguistic annotation infrastucture: FoLiA, its tools, and FLAT

Maarten Van Gompel

Attribute Obfuscation with Gradient Reversal

Chris Emmery, Enrique Manjavacas and Grzegorz Chrupała

Authorship Attribution on the Dark Web

Maya Sappelli and Stephan Raaijmakers

Automated detection of offensive language on social media

Finn Rübsaam, Roser Morante and Fleur van der Houwen

Automatic Detection and Correction of dt-mistakes Using Recurrent Neural Networks

Yannick Laevaert, Geert Heyman and Marie-Francine Moens

Automatically grading students' answers to exam questions

Johanna de Vos

Automatic image description for Dutch

Emiel van Miltenburg

Automating Coherence Analysis for the Computationally Challenged

Erwin R. Komen and Jet Hoek

Children are more likely to store coherent rather than frequent under-segmented chunks: Quantitative evidence from a corpus study

Robert Grimm, Steven Gillis and Walter Daelemans

Classifying Abnormalities in Mammography Reports in Dutch

Tim Loyen and Roser Morante

Clinical Machine Comprehension with Case Reports

Simon Suster and Walter Daelemans

Comparing MT errors in NMT, SMT and RBMT output for English-Dutch

Laura Van Brussel, Arda Tezcan and Lieve Macken

 $Computational\ linguistic\ opportunities\ at\ the\ Koninklijke\ Bibliotheek$

Martijn Kleppe, Lotte Wilms and Steven Claeyssens

Content and Annotation Compatibility Across Corpora: The Case of Events

Chantal van Son, Lora Aroyo, Oana Inel, Roser Morante and Piek Vossen

Dialect mapped: knowledge rich or knowledge poor?

Hans van Halteren, Roeland van Hout and Romy Roumans

Evaluation of Different Dialogue Corpora for Training a Chatbot

Elizaveta Kuzmenko

Exploring automated template generation for data-to-text systems

Chris van der Lee, Emiel Krahmer and Sander Wubben

From distant to close reading: evaluating the use of LDA topic modeling

Christoph Aurnhammer, Iris Cuppen, Inge van de Ven and Menno van Zaanen

Gold Standard for Multilingual Automatic Term Extraction from Comparable Corpora

Ayla Rigouts Terryn, Véronique Hoste and Els Lefever

Improving Sense Disambiguation of Potentially Idiomatic Expressions

Hessel Haagsma

Language change in the Corpus Gesproken Nederlands: An apparent-time study of verb cluster order Marieke Olthof, Maud Westendorp and Jelke Bloem

Lexical preferences in the Perfect construction

Martijn van der Klis

Modeling relations in a referential game

Lieke Gelderloos, Afra Alishahi, Grzegorz Chrupała and Raquel Fernández

Named Entity Recognition in Dutch Archaeological Research Reports
Alex Brandsen, Suzan Verberne and Milco Wansleeben

Ochre, a Toolbox for OCR Post-Correction

Janneke Van Der Zwaan and Lotte Wilms

Predicting the asking price of real estate offers

Hendrik Prins and Malvina Nissim

Reddit as a Corpus for Data-Driven Dialogue Modelling Bram Willemsen

Semi-Automated Anonymization of Dutch Court Cases Simon Brugman and Arjen P. de Vries

Sentiment Analysis: Skipping Mumbo Jumbo and Revealing How We Actually Do It!!

Md Ataur Rahman and Rudra Pratap Deb Nath

Spelling variation in Dutch tweets

Hans Van Halteren and Nelleke Oostdijk

Stylometric Text Analysis for Adolescents with Autism Spectrum Disorder Luna De Bruyne and Walter Daelemans

Syntactic properties of skip-thought vectors

Bart Broere

TICCLAT: Text-Induced Corpus Correction and Lexical Assessment Tool
Martin Reynaert

The Circumstantial Event Ontology (CEO): A New Resource for Tracking Event Chains based on Implicit Semantic Relations.

Roxane Segers, Piek Vossen and Tommaso Caselli

The Claus Case: Exploring the Use of Propositional Idea Density for Alzheimer Detection Silke Marckx, Ben Verhoeven and Walter Daelemans

Towards an automatic identification of tone of voice in online conversations Christine Liebrecht and Charlotte Van Hooijdonk

Using Location and Dialect to Classify Twitter Users from the Netherlands and Flanders Ali Hürriyetoğlu and Piet Daas

/instituut voor de Nederlandse taal/

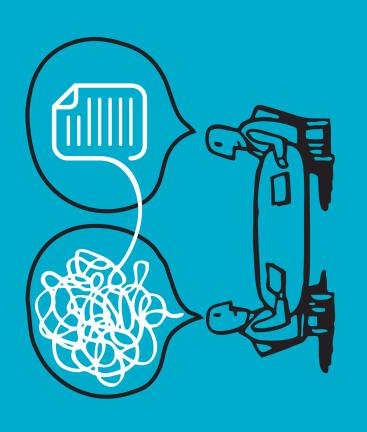
The *Instituut voor de Nederlandse Taal* (or Dutch Language
Institute) is *the* place for anyone
who wants to know anything
about the Dutch language through
the centuries.

It is a scholarly, but at the same time easily accessible institute that studies all aspects of the Dutch language, including its vocabulary, grammar and linguistic variations.

The institute collects old and new Dutch words, updates important reference works such as the *Algemene Nederlandse Spraakkunst*, the main standard work on Dutch grammar, and creates terminology lists to make professional jargons accessible.

The Dutch Language Institute...

- takes part in IMPACT
- participates in CLARIAH
- is a certified CLARIN B
 Centre
- provides GiGaNT, a computational lexicon of 15 centuries of Dutch, and various other lexica
- manages a large number of Dutch text and speech corpora
- develops tools for linguistic research, such as a tool for named entity recognition
- provides BlackLab an opensource corpus retrieval engine that allows fast, complex searches with accurate hit – highlighting on large, tagged and annotated, bodies of text



(↑ Our Talent Team rotated this on purpose)

textkernel

Innovation, Artificial Intelligence and Machine Learning for matching jobs and people

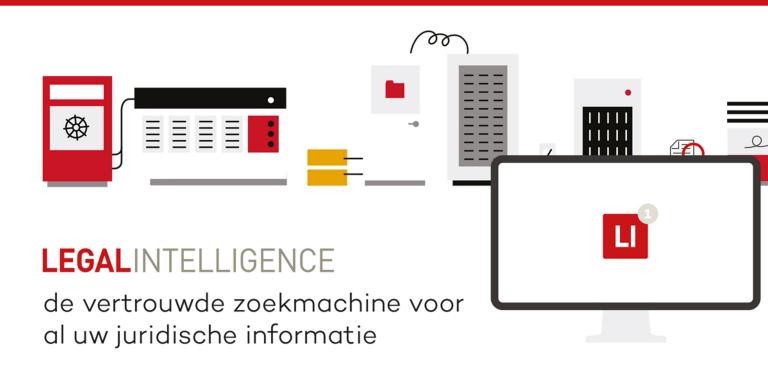
- · Information extraction and retrieval
- · Recommendation systems
- Deep Learning and Big Data
- Week long hackathon
- · Influence on products

We're hiring!

Visit textkernel.careers or email us at jobs@textkernel.com

LEGALINTELLIGENCE

Beter zoeken. Zeker weten.





Notes	

Notes	

Notes	

