CORE Skills

Delivering Data Science
In Resources & Energy

NTLP

Knowledge Discovery from Natural Language Processing & Text Mining

DAY 12: Special Data Types

15-Day Data Science Springboard

Assoc Prof Wei Liu & Tyler Bikaun, Department of Computer Science & Software Engineering, UWA
wei.liu@uwa.edu.au
tyler.bikaun@research.uwa.edu.au

Program partners

_dm

THE UNIVERSITY OF WESTERN AUSTRALIA

Curtin University

A CORE partnership

Program Timeline

DAY 12: Special Data Types – Natural Language Processing & Text Mining

2-hour Leading Data Scientists Leader Support	Preparatory		Introduction to Data Projects	Data Analysis			Data & Communication Sandbox	Data Fusion and Machine Learning		Data Fusion Sandbox	Special Data Types - Time-series & Networks	Special Data Types - Knowledge Discovery from Natural Language Processing & Text Mining	Special Data	Capstone Project Development & Presentation	Capstone Propeller
2-hours	Day 1	Day 2	Day 3	Day 4	Day 5	Day 6	Day7	Day 8	Day 9	Day 10	Day 11	Day 12	Day 13	Day 14	Day 15
Enabling your people's data science upskilling & project delivery in 15 day program	Introduction to the program tools	Introduction to the program tools	Zero to Data Science in a day	Getting to know the Program Tools: Data munging and exploratory data analysis	Simple predictions: Regression and statistical model building	Multivariate analysis and model building	Effective data storytelling: Communicating results to non-technical audiences	Pros and cons of commonly used statistical and machine learning techniques I	Pros and cons of commonly used statistical and machine learning techniques II	Consolidate approaches covered and test on datasets	The 4th dimension and predictions	Finding needles in wordstacks	Spatial analytics and predictions	Capstone Project pitches to leadership	Project Review Day

Schedule

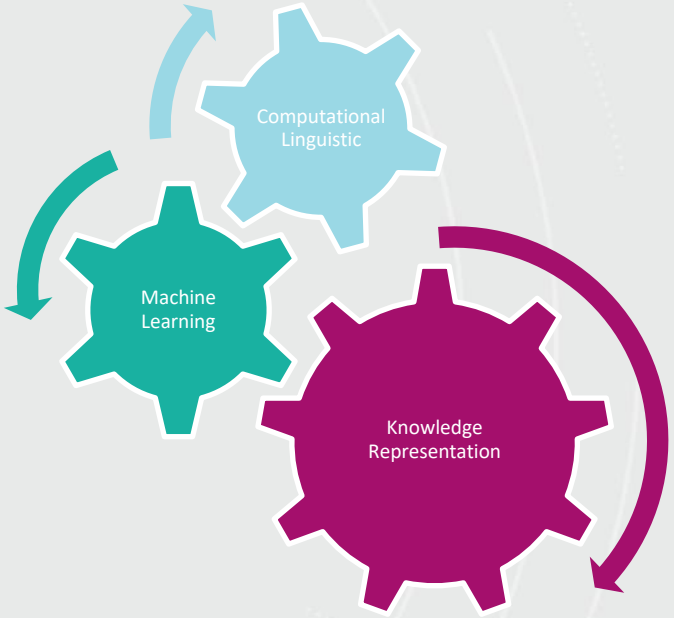
DAY 12

AWST	AEST	Agenda	
08:00	10:00	Open JupyterHub, Q&A	
08:15	**10:15**	**Part 1: NLP Fundamentals** (*5 min break ~8:45/10:45)*	Tyler/Wei
09:30	*11:30*	*Morning Tea*	
09:45	**11:45**	**Part 2: Language Representation** (*5 min break ~10:30/12:30)*	Tyler
11:15	*13:15*	*Lunch*	
12:00	**14:00**	**Part 3: Supervised NLP and Annotation** (*5 min break ~12:45/14:45)*	Tyler
13:30	*15:30*	*Afternoon Tea*	
13:45	**15:45**	Project Update	Tamryn
14:00	**16:00**	**Part 4: Unsupervised NLP** (*5 min break ~14:30/16:30)*	Wei
15:15	17:15	Q&A, Reflections, Takeaways, Menti Feedback	Tamryn
15:30	17:30	Close	

A day on Natural Language Processing and Text Mining

What is today about?

- Natural Language Processing
- Language Representation
- Supervised Learning
- Natural Language Annotation
- Unsupervised Learning



Associate Prof. Wei Liu & Tyler Bikaun, UWA

Aims & Learning Outcomes – Day 12

Aims

1. Understand practical strategies and applications for using text
2. Provide a foundation for NLP fundamentals including text wrangling, pre-processing, and representation
3. Gain familiarity with supervised and unsupervised learning
4. Understand the natural language annotation process

Learning Outcomes

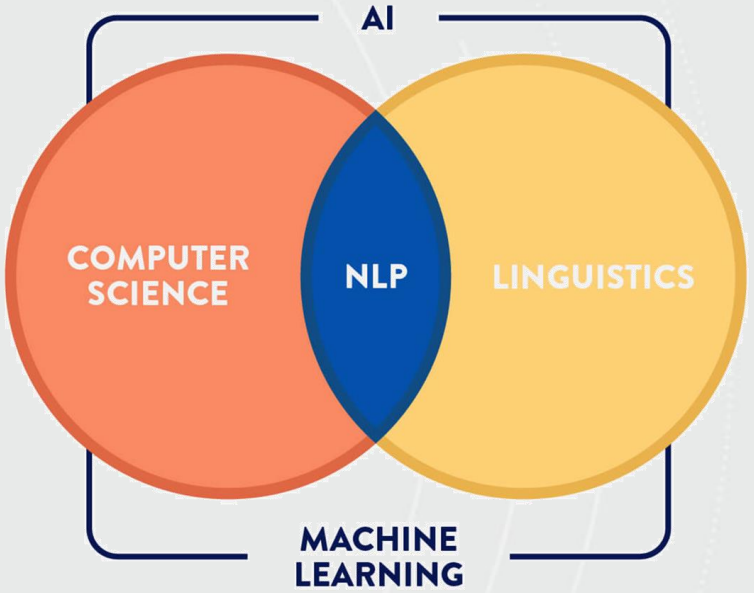
1. Understand how to wrangle, pre-process and gain insight into text
2. Perform a range of supervised learning tasks
3. Understand the text annotation process
4. Perform unsupervised learning


12.0 Overview
What can we get out of natural language texts

Tasks in Natural Language Processing[1]

NLP Tasks

- Data-to-Text Generation
- Grammatical error correction
- Lexical normalization
- Machine translation
- Named entity recognition
- Question answering
- Relation extraction
- Sentiment Analysis
- Simplification
- Summarization
- Text classification
- …

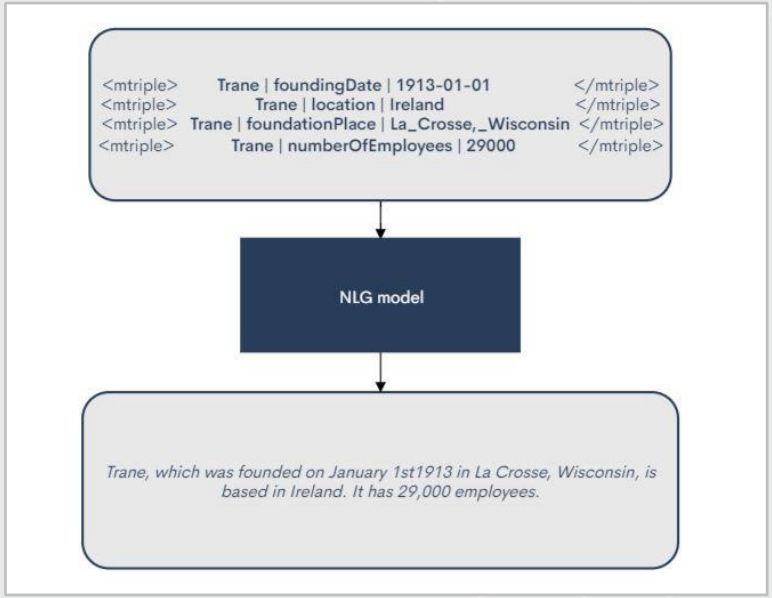


1 - http://nlpprogress.com/

Tasks in Natural Language Processing[1]

NLP Tasks

- **Data-to-Text Generation**
- Grammatical error correction
- Lexical normalization
- Machine translation
- Named entity recognition
- Question answering
- Relation extraction
- Sentiment Analysis
- Simplification
- Summarization
- Text classification
- ...



1 - http://nlpprogress.com/

Tasks in Natural Language Processing[1]

NLP Tasks

- Data-to-Text Generation
- **Grammatical error correction**
- Lexical normalization
- Machine translation
- Named entity recognition
- Question answering
- Relation extraction
- Sentiment Analysis
- Simplification
- Summarization
- Text classification
- ...

Input (Erroneous)	Output (Corrected)
A important part of my life have been a people that stood by me.	An important part of my life has been the people who stood by me.

1 - http://nlpprogress.com/

Tasks in Natural Language Processing[1]

NLP Tasks

- Data-to-Text Generation
- Grammatical error correction
- **Lexical normalization**
- Machine translation
- Named entity recognition
- Question answering
- Relation extraction
- Sentiment Analysis
- Simplification
- Summarization
- Text classification
- ...

source: got **exo** to share, **u** interested? Concert in **hk** !

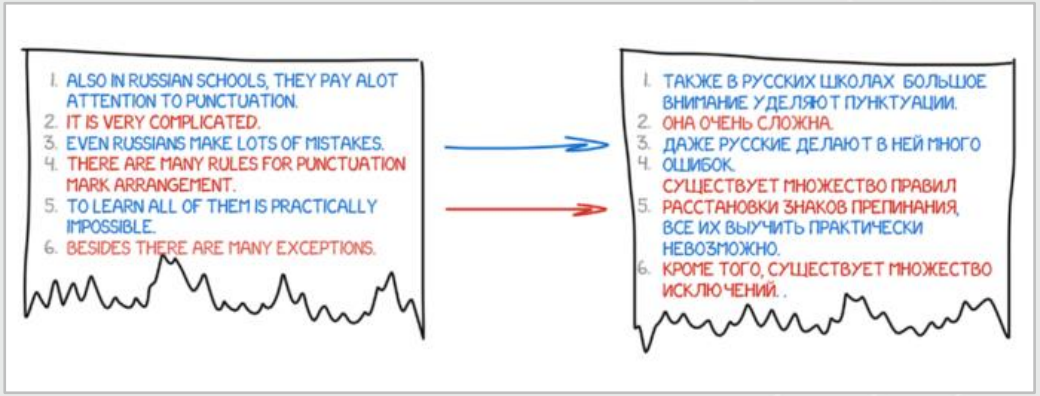
target: got **extra** to share, **are you** interested? Concert in **hong kong** !

1 - http://nlpprogress.com/

Tasks in Natural Language Processing[1]

NLP Tasks

- Data-to-Text Generation
- Grammatical error correction
- Lexical normalization
- **Machine translation**
- Named entity recognition
- Question answering
- Relation extraction
- Sentiment Analysis
- Simplification
- Summarization
- Text classification
- ...

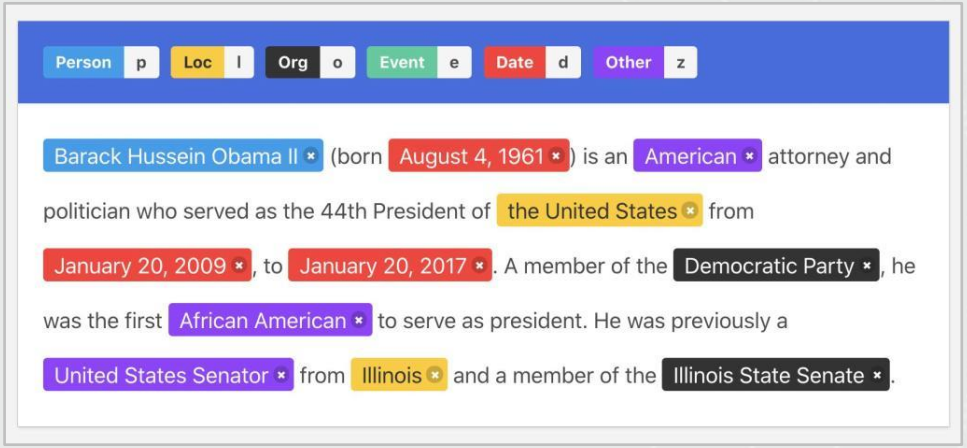


1 - http://nlpprogress.com/

Tasks in Natural Language Processing[1]

NLP Tasks

- Data-to-Text Generation
- Grammatical error correction
- Lexical normalization
- Machine translation
- **Named entity recognition**
- Question answering
- Relation extraction
- Sentiment Analysis
- Simplification
- Summarization
- Text classification
- ...

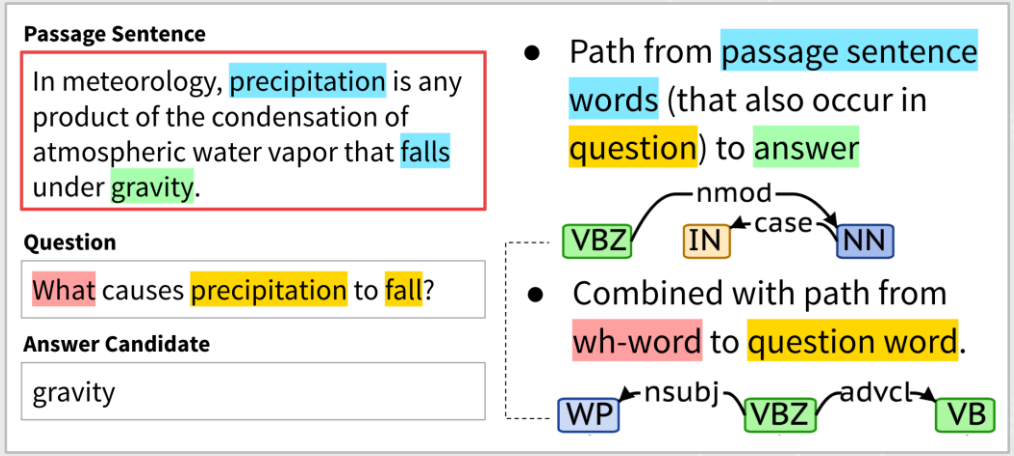


1 - http://nlpprogress.com/

Tasks in Natural Language Processing[1]

NLP Tasks

- Data-to-Text Generation
- Grammatical error correction
- Lexical normalization
- Machine translation
- Named entity recognition
- **Question answering**
- Relation extraction
- Sentiment Analysis
- Simplification
- Summarization
- Text classification
- ...

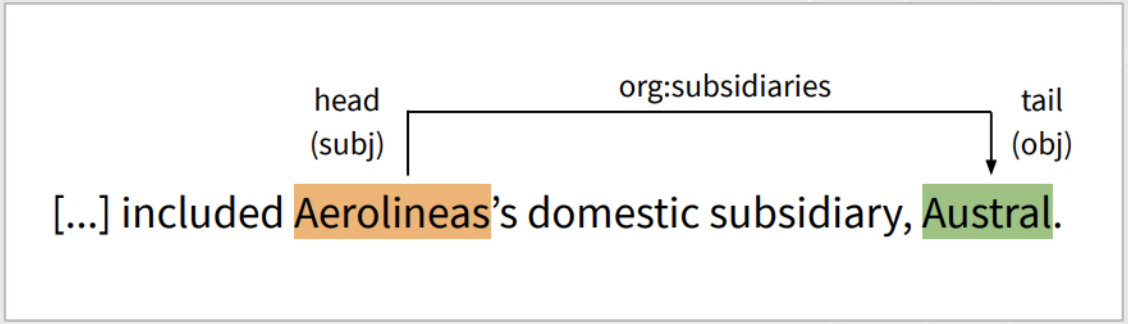


1 - http://nlpprogress.com/

Tasks in Natural Language Processing[1]

NLP Tasks

- Data-to-Text Generation
- Grammatical error correction
- Lexical normalization
- Machine translation
- Named entity recognition
- Question answering
- **Relation extraction**
- Sentiment Analysis
- Simplification
- Summarization
- Text classification
- ...

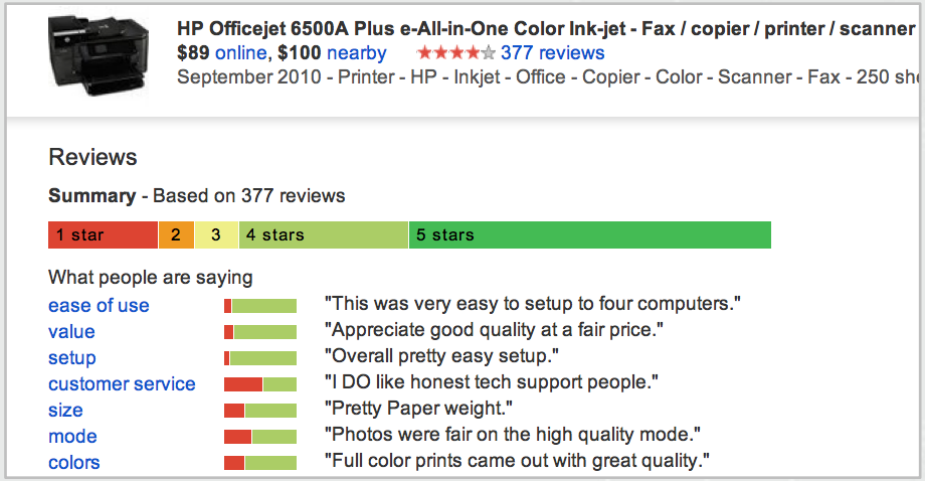


1 - http://nlpprogress.com/

Tasks in Natural Language Processing[1]

NLP Tasks

- Data-to-Text Generation
- Grammatical error correction
- Lexical normalization
- Machine translation
- Named entity recognition
- Question answering
- Relation extraction
- **Sentiment Analysis**
- Simplification
- Summarization
- Text classification
- ...



1 - http://nlpprogress.com/

Tasks in Natural Language Processing[1]

NLP Tasks

- Data-to-Text Generation
- Grammatical error correction
- Lexical normalization
- Machine translation
- Named entity recognition
- Question answering
- Relation extraction
- Sentiment Analysis
- **Simplification**
- Summarization
- Text classification
- ...

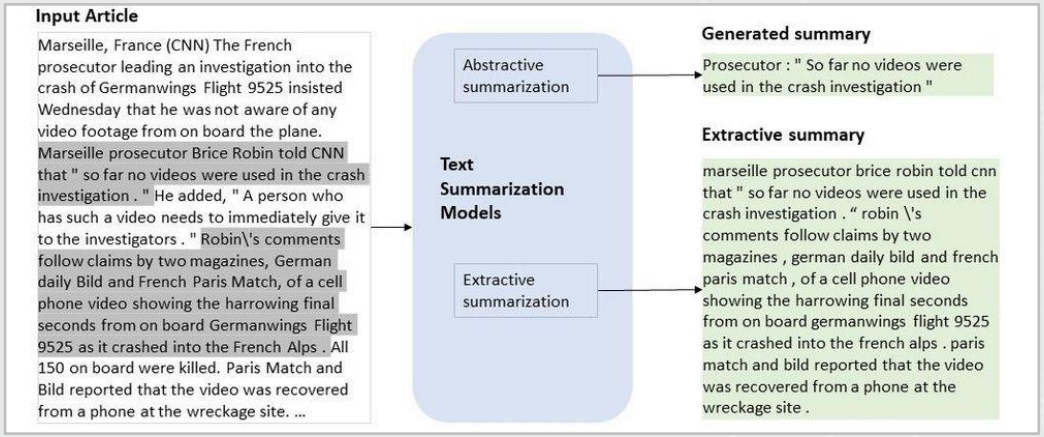
Example of a Complex Sentence	Example of a Simplified Sentence
Grammarly provides assistance in order to optimize users' communication.	Grammarly helps people communicate.

1 - http://nlpprogress.com/

Tasks in Natural Language Processing[1]

NLP Tasks

- Data-to-Text Generation
- Grammatical error correction
- Lexical normalization
- Machine translation
- Named entity recognition
- Question answering
- Relation extraction
- Sentiment Analysis
- Simplification
- **Summarization**
- Text classification
- ...

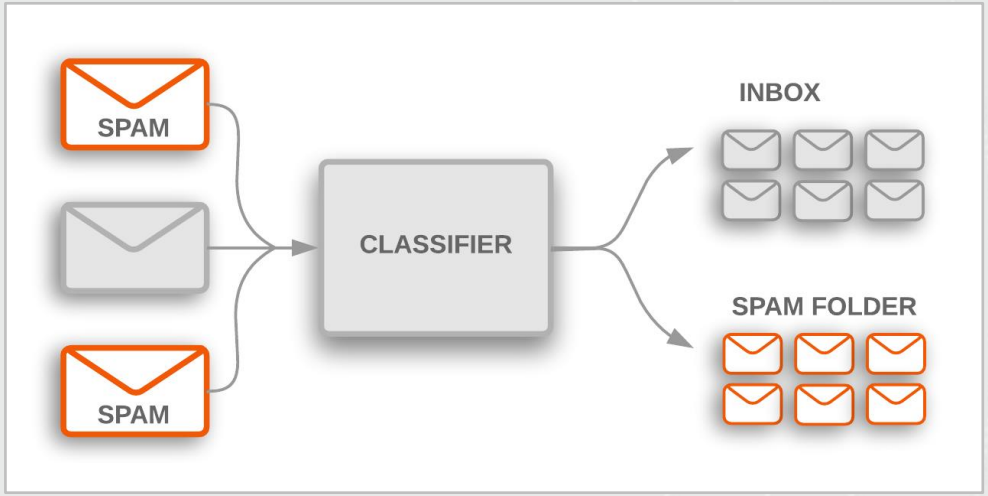


1 - http://nlpprogress.com/

Tasks in Natural Language Processing[1]

NLP Tasks

- Data-to-Text Generation
- Grammatical error correction
- Lexical normalization
- Machine translation
- Named entity recognition
- Question answering
- Relation extraction
- Sentiment Analysis
- Simplification
- Summarization
- **Text classification**
- ...



1 - http://nlpprogress.com/

Our Work and Projects

Programming Language or Natural Language?

Natural Language

- General
- Grammatically Correct
- Large Vocabulary
- Not noisy
- Examples: Wikipedia

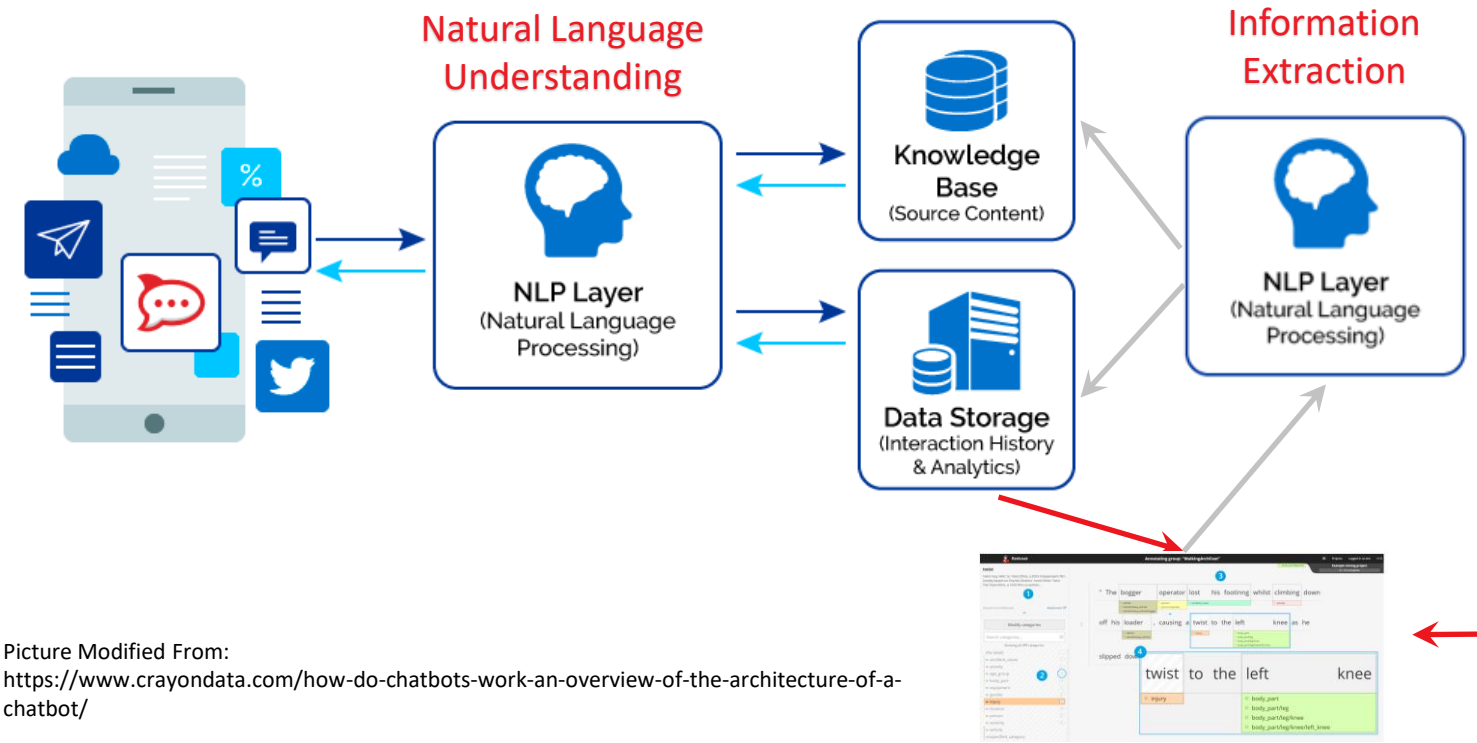
Technical Language

- ✓ Domain Specific
- ✓ Small Vocabulary
- × Terse
- × Jargons, Misspelling
- Examples: Safety logs, Maintenance work orders, geological reports

Technical language processing (TLP) aims to understand **technical languages**, a subset of natural languages that appears in industry-specific contexts.

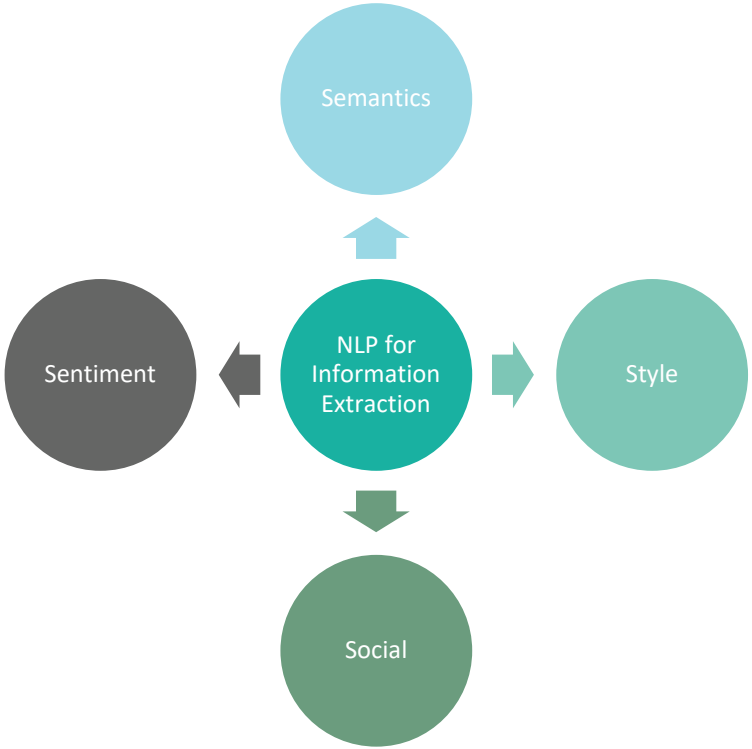
- Lexical Normalisation
- Grammatical Error Correction
- No pretrained embeddings
- No ground truth
- Low-resource datasets

Language – Natural or Technical?



Picture Modified From:
https://www.crayondata.com/how-do-chatbots-work-an-overview-of-the-architecture-of-a-chatbot/

What can we extract out of text? – S4



What, When, Where, How and Why?

Semantics (Entities)
- Entity Recognition

Social (Relations)
- Relation Extraction
- Triple Extraction (Knowledge Graphs)

Style
- Stylometry (fraud, authorship, forensics)
- Grammar (Industry Synthetic Data Generation)

Sentiment
- Social Reviews (e.g. Fraud Detection)

Picture Credit: https://research.aimultiple.com/chatbot-architecture/

The hidden treasure in your dataset

	W	X	Y	Z	AA	AB	AC	AD	AE	AF	AG	
1	Place Of Injury		Activity	Injury Typ	Body Part	Nature Of	Occupatio	Work Type	Company	Contractor Company		Detailed Description
2	0471 - Processing Pla		8230 - Rep	2510 - C/B	6300 - Hea	5600 - Cru	719000 - E	Full Time	Contracto	Contractor		Traumatic crush injury to the head and chest.
3	0610 - W.Shop Heavy		8130 - Rep	2510 - C/B	6500 - Cor	5400 - Fra	631000 - F	Full Time	Contracto	Contractor		Two fitters were undertaking maintenance work when an accident occurred resulting i
4	0451 - Tailings Storag		7900 - Mis	2130 - S/B	6920 - Mu	5600 - Cru	411000 - P	Full Time	Contracto	Contractor		At approximately 9.20am on Wednesday 4th December a contractor working on the ta
5	0610 - W.Shop Heavy		2110 - Wa	1210 - Fall	5300 - Kne	5200 - Stra	631000 - F	Full Time	Contracto	Contractor		Drill fitter had finished planned services for the shift and had returned the tools to the
6	0370 - Storage Yard/		2450 - Mo	9100 - Bite	5400 - Low	6310 - Bite	359000 - N	Full Time	Contracto	Contractor 1		The injured person identified a stinging sensation on his left shin area. The injured pers
7	0471 - Processing Pla		7900 - Mis	9100 - Bite	5200 - Upp	6310 - Bite	359000 - N	Full Time	Contracto	Contractor 1		Injured person was sleeping when he felt a stinging sensation on his right thigh area. T
8	0464 - Reagent / Raw		2450 - Mo	2510 - C/B	4630 - Fin	6100 - Lac	411000 - P	Full Time	Contracto	Contractor 10		An Operator was transporting signs in a loader bucket. Whilst removing one of the sign
9	0464 - Reagent / Raw		2320 - Lift	4200 - O.X	3100 - Bac	5200 - Stra	372000 - C	Full Time	Contracto	Contractor 10		On 11/11/13 at approximately 0900hrs employee was assisting two other operators in
10	0128 - Level (Develop		2410 - Pull	1120 - Fall	4620 - Har	5200 - Stra	241000 - E	Full Time	Contracto	Contractor 11		Operator was climbing onto the rear of the bogger to straighten a hand rail when it gav
11	0128 - Level (Develop		2180 - Get	3150 - Ste	5500 - Ank	5200 - Stra	269000 - L	Full Time	Contracto	Contractor 11		Service crew member stepped down from work cage onto uneven ground and rolled le
12	0950 - Car Park		2110 - Wa	3120 - Ste	5500 - Ank	5200 - Stra	165000 - S	Full Time	Contracto	Contractor 11		walking in car park from office to workshop & rolled ankle
13	0128 - Level (Develop		2180 - Get	1420 - Fall	5300 - Kne	5200 - Stra	241000 - E	Full Time	Contracto	Contractor 11		The bogger operator lost his footinng whilst climbing down off his loader, causing a twi
14	0130 - Underground		2110 - Wa	1232 - Fall	3181 - Bac	6800 - Pai	241000 - E	Full Time	Contracto	Contractor 11		Person was walking around the bogger in the workshop and trod in some grease that v

What, When, Where, How and Why?

- Individual vs. Population
- Query vs. Machine Learning

activity location location

walking in car park from office to workshop

and rolled ankle

location injury body part

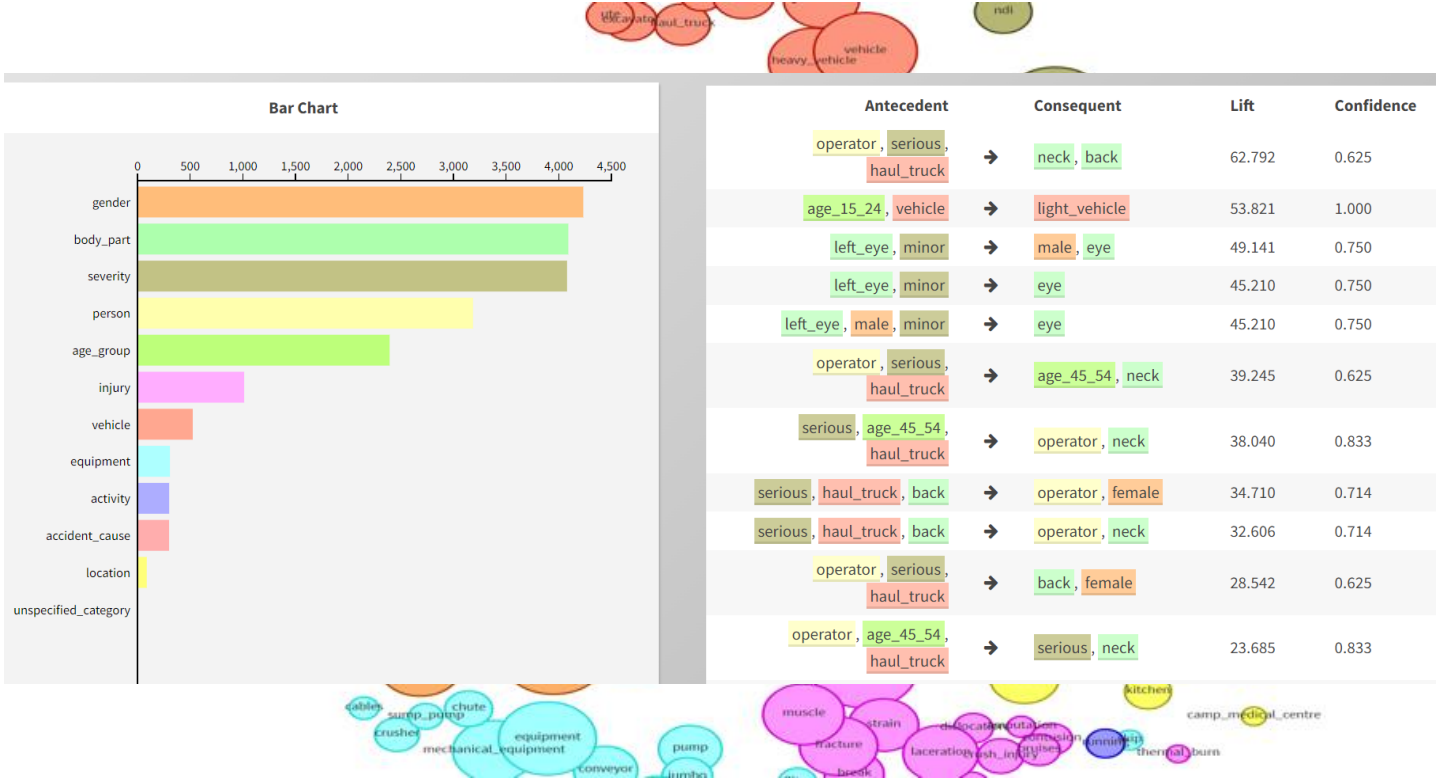
NTLP

S4 – Semantics and Social

Semantics

NLP for Information Extraction

Social

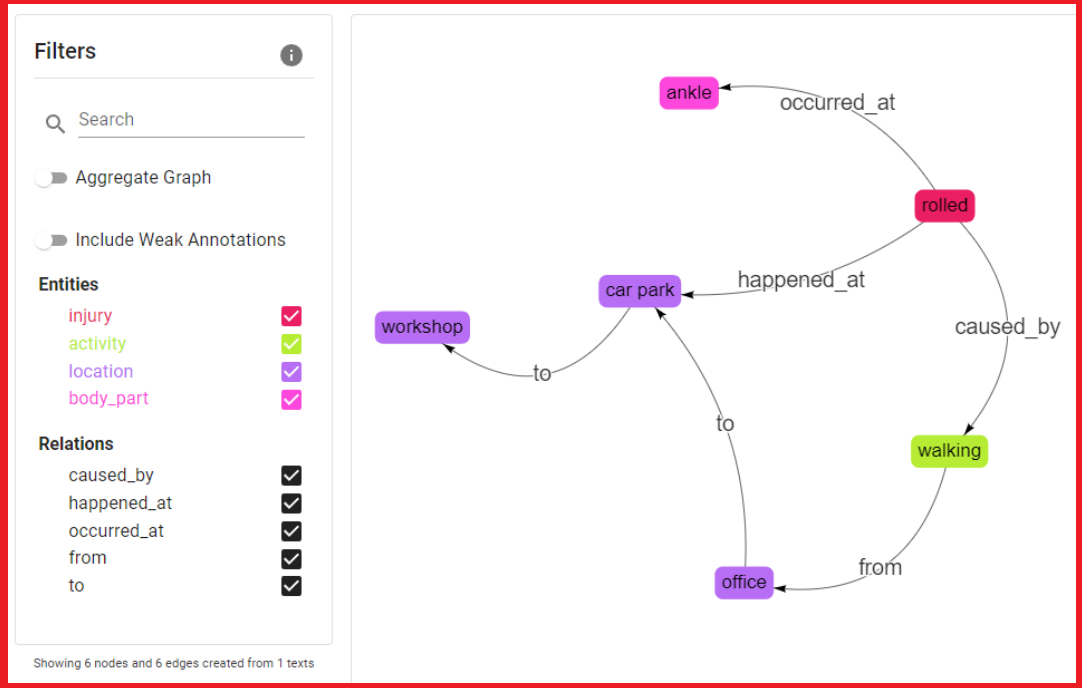


Antecedent		Consequent	Lift	Confidence
operator, serious, haul_truck	→	neck, back	62.792	0.625
age_15_24, vehicle	→	light_vehicle	53.821	1.000
left_eye, minor	→	male, eye	49.141	0.750
left_eye, minor	→	eye	45.210	0.750
left_eye, male, minor	→	eye	45.210	0.750
operator, serious, haul_truck	→	age_45_54, neck	39.245	0.625
serious, age_45_54, haul_truck	→	operator, neck	38.040	0.833
serious, haul_truck, back	→	operator, female	34.710	0.714
serious, haul_truck, back	→	operator, neck	32.606	0.714
operator, serious, haul_truck	→	back, female	28.542	0.625
operator, age_45_54, haul_truck	→	serious, neck	23.685	0.833

Software Demo at: https://nlp-tlp.org/aquila/
Developed by Dr. Michael Stewart

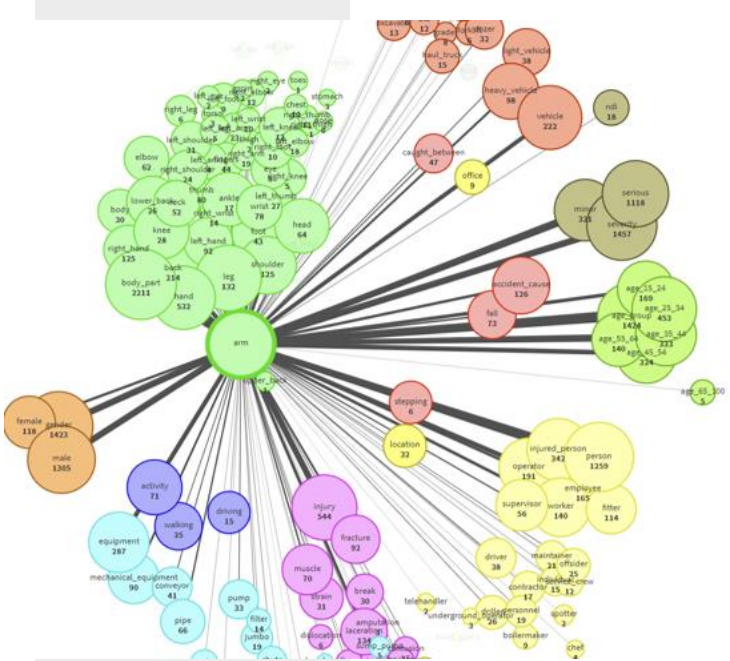
Annotation Tools Developed at UWA – QuickGraph and Lexiclean

Maintenance Work Order Processing - ARC Training Centre for Transforming Maintenance through Data Science



Demos Available at: https://nlp-tlp.org/
By PhD Student: Tyler Bikaun

Award Winning NLP & KG Solutions

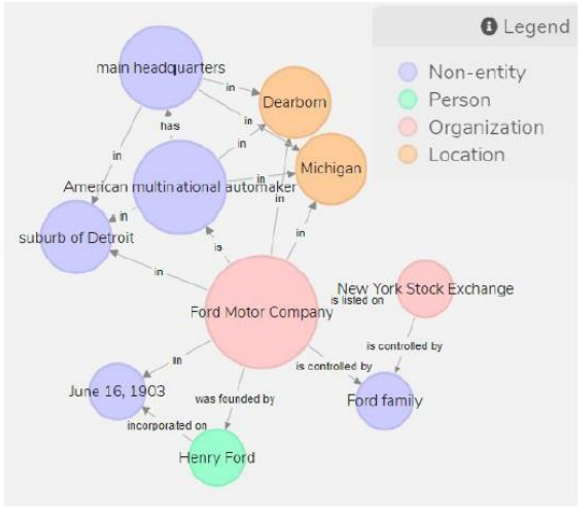


WAITTA INCITE Award 2019

GAMEs
GREEN ASSET MANAGEMENT & EVALUATION SYSTEM

WAITTA INCITE Award Finalist 2022

Ford Motor Company is an American multinational automaker that has its main headquarters in Dearborn, Michigan, a suburb of Detroit. The company was founded by Henry Ford and incorporated on June 16, 1903. The company is listed on the New York Stock Exchange and it is controlled by the Ford family.



ICDM Knowledge Graph Contest 1st Prize

TECHNICAL LANGUAGE PROCESSING | TEXT TO KNOWLEDGE GRAPHS | OUR TOOLS & SOLUTIONS | HUMAN-DIGITAL WORKFORCE

UWA Natural Language and Technical Language Processing Group

https://nlp-tlp.org

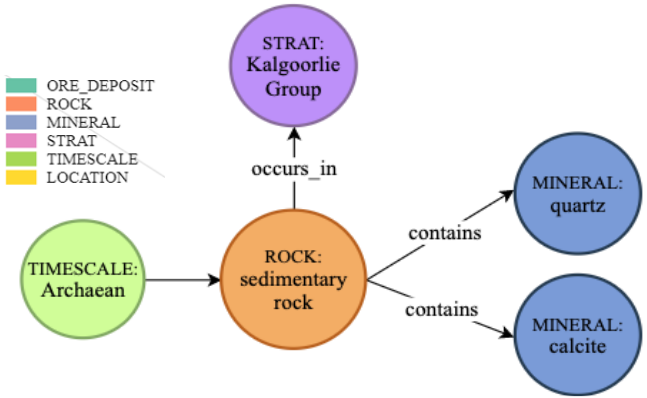
KG Extraction from Geological Surveys

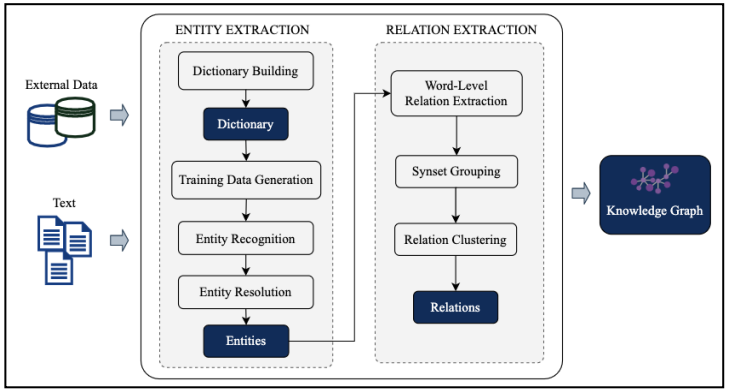
PhD Thesis of Majigsuren Enkhsaikhan
Supported by GSWA

A. Archaean sedimentary rocks occurred within the Kalgoorlie Group. Most sedimentary rocks contain either quartz or calcite.

B.
Archaean sedimentary rocks —> sedimentary rocks
sedimentary rocks —occurred within—> Kalgoorlie Group
sedimentary rocks —contain—> quartz
sedimentary rocks —contain—> calcite

C.





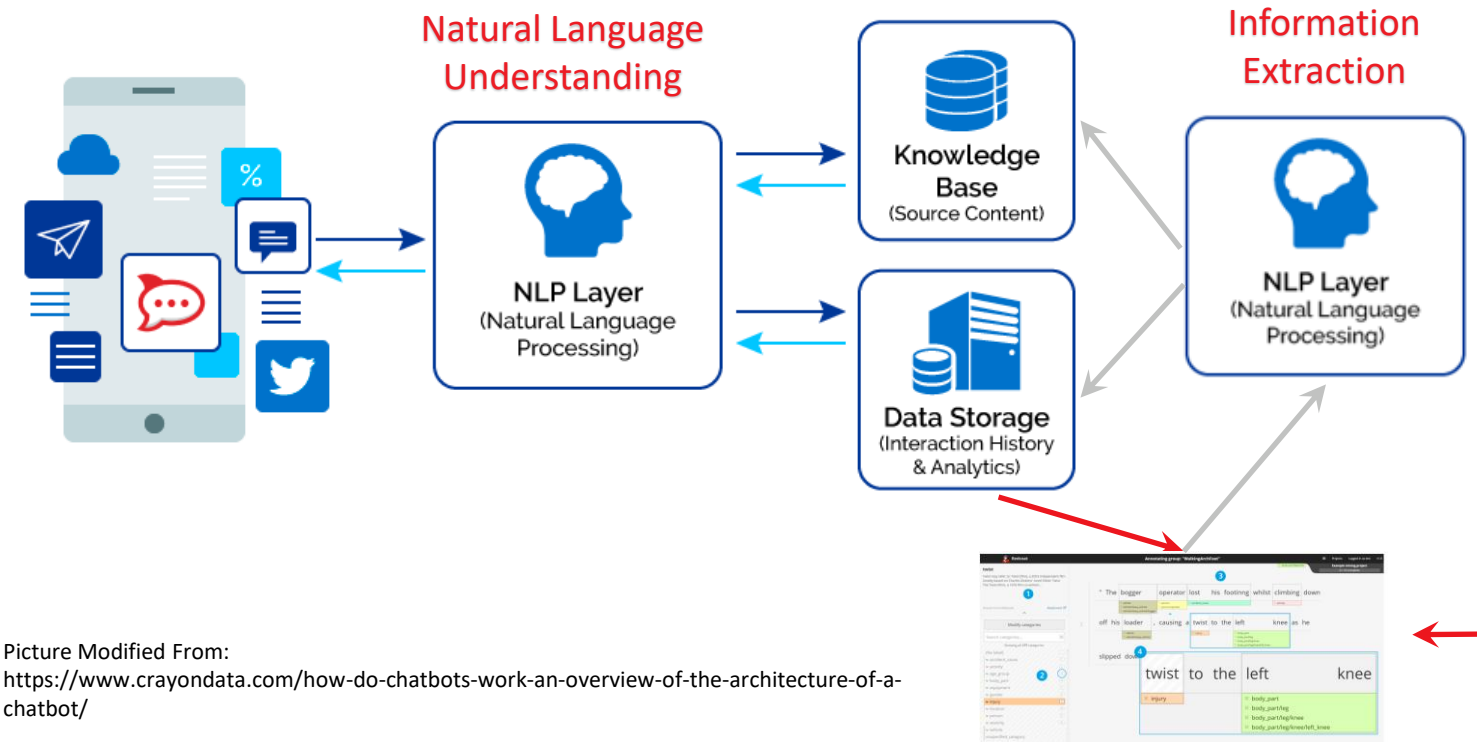
Knowledge Graph Schema

Enkhsaikhan et al. (2021), "Understanding Mineral Depositional Conditions using Machine Reading of Text", ***Ore Geology Reviews***

Enkhsaikhan et al. (2020), "Auto-Labelling Entities in Low Resource Text: A Geological Case Study", ***Knowledge and Information Systems***

Enkhsaikhan et al. (2018), "Towards geological knowledge discovery using vector-based semantic similarity", ***Proceedings of the 14th Int. Conf. on Advanced Data Mining and Applications***

TLP – redefining the workforce



Picture Modified From:
https://www.crayondata.com/how-do-chatbots-work-an-overview-of-the-architecture-of-a-chatbot/

Human – Digital Workforce: Trends in Automation 2022

People will work side-by-side with virtual robotic assistants—sharing tasks, passing work back and forth, collaborating.

Within the next 5 years, hybrid human-digital workforces will

- Work in a task-oriented (rather than software application-oriented) manner, and
- Be part of virtual assembly line

THANK YOU

NLP-TLP Group: https://nlp-tlp.org/

Academics

Wei Liu
Associate Professor
UWA Profile | LinkedIn

Melinda Hodkiewicz
Professor
UWA Profile | LinkedIn

Tim French
Senior Lecturer
UWA Profile | LinkedIn

Michael Stewart
Postdoctoral Research Fellow
UWA Profile | LinkedIn

Eun-Jung Holden
Professor
UWA Profile | LinkedIn

Caren Han
Lecturer
UWA Profile

Naeha Sharif
Lecturer
UWA Profile | LinkedIn

PhD Students

Ziyu Zhao
PhD Student
An Efficient Neural Probabilistic Logical Resolution for Multi-class Multi-label Entity Typing
LinkedIn

Tyler Bikaun
PhD Student
Technical Language Processing for Industrial Maintenance Records
LinkedIn

Chau Nguyen Duc Minh
PhD Candidate
Query Embedding for Long Reasoning over Natural-Technical Domains

Caitlin Woods
PhD Student
Adaptive User Interfaces for Industrial Maintenance Procedures
Website | LinkedIn

Tom Smoker
PhD Student
Rectifying knowledge graph link prediction using embedding-enhanced ontologies
LinkedIn

AI 2022: https://ajcai2022.org/

wei.liu@uwa.edu.au
linkedin.com/in/wei-liu-b0b4521ab


12.1 Fundamentals of NLP
Text wrangling, pre-processing and analysis

Fundamentals of Natural Language Processing

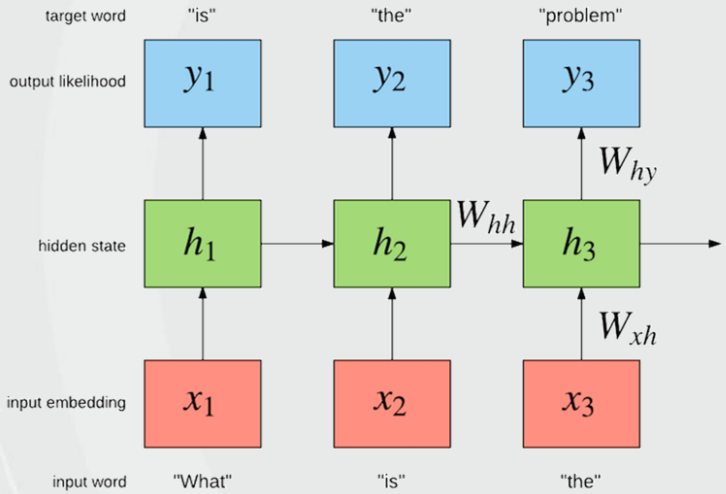
Agenda (Hands-on Session) - An introduction to the Natural Language Tool Kit (NLTK)

- Tokenization
- Stop-words
- Feature distributions
- N-grams
- Stemming
- Concordancing
- Dispersion plotting
- Bi-gram significance
- Word contexts
- Word similarities
- Parts-of-Speech (POS)
- Named entity recognition


12.2 Fundamentals of NLP
Representing natural language

Word Embeddings: One-hot Vectors

Word representations – one-hot vectors



(Socher, 2018)

What is the name of the prime minister of Australia?

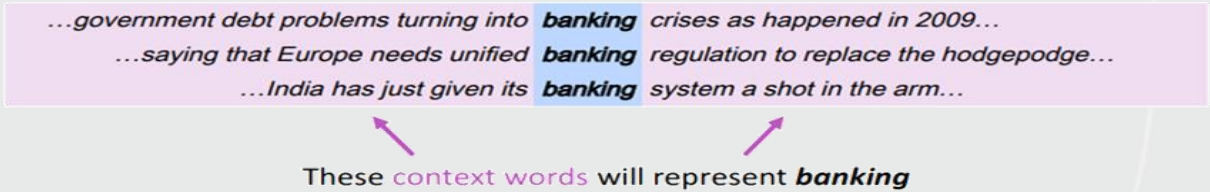
What	=	[10000000]
is	=	[01000000]
the	=	[00100000]
name	=	[00010000]
of	=	[00001000]
the	=	[00100000]
prime	=	[00000100]
minister	=	[00000010]
of	=	[00001000]
Australia	=	[00000001]

Word Embeddings: Learning Context

Representing words by their context

You shall know a word by the company it keeps

J. R. Firth 1957



prime	=	[0,0,0,0,0,1,0,0]	→	prime	=	[-0.1, 0.1, 0.2, 0.3, 0.3, 0.4, 0.1, 0.7]
minister	=	[0,0,0,0,0,0,1,0]		minister	=	[-0.2, 0.2, 0.4, 0.2, 0.2, 0.2, 0.1, 0.6]

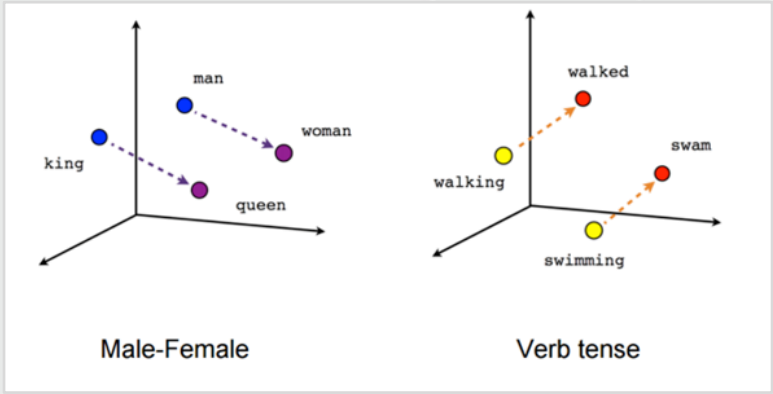
One-hot **Embeddings**

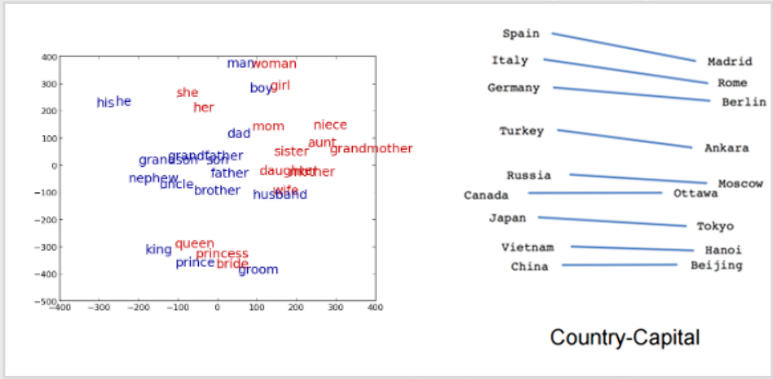
Word Embeddings: Power of Embeddings

Imagine doing simple math with words

"A is to B as C is to D" tasks

$-\text{word}_A - \text{word}_B - \text{word}_C \cong \text{word}$





http://rare-technologies.com/word2vec-tutorial/

Word Embeddings: Analogy Task Examples

“A is to B as C is to D” tasks

groom bride husband wife
groom bride king queen
groom bride man woman
groom bride nephew niece
groom bride policeman policewoman
groom bride prince princess

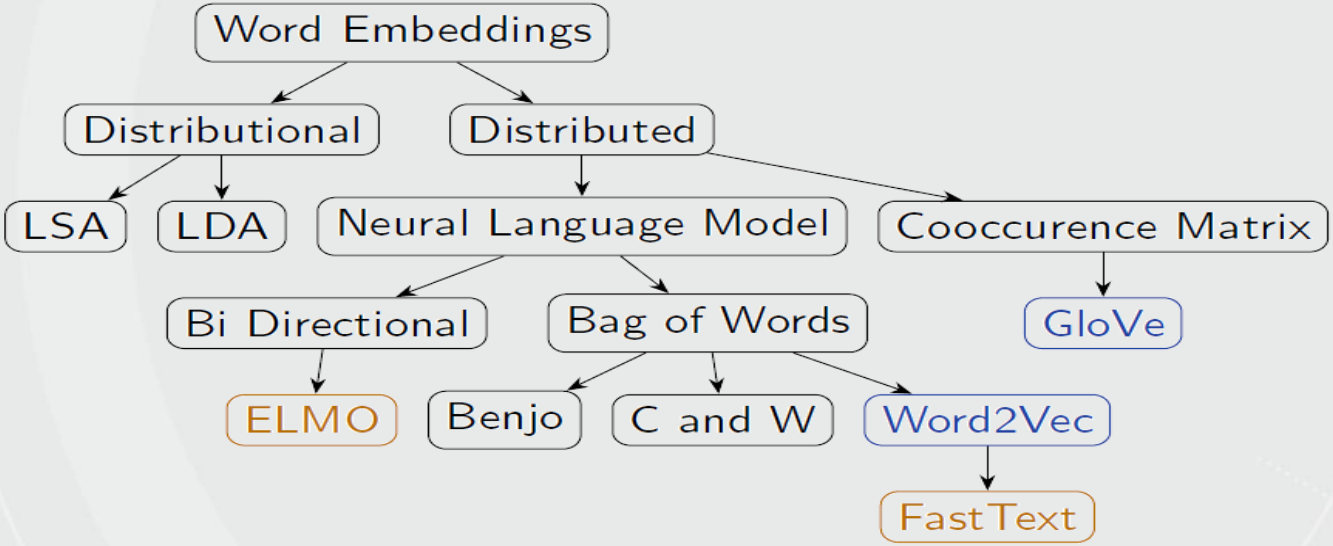
Athens Greece Baghdad Iraq
Athens Greece Bangkok Thailand
Athens Greece Beijing China
Athens Greece Berlin Germany
Athens Greece Bern Switzerland

short shorter small smaller
short shorter smart smarter
short shorter strong stronger
short shorter tall taller
short shorter tight tighter
short shorter tough tougher
short shorter warm warmer
short shorter weak weaker
short shorter wide wider
short shorter young younger
short shorter bad worse

Associate Prof. Wei Liu & Tyler Bikaun, UWA

Word Embeddings: Techniques

Many ways of obtaining word vectors



New techniques are being developed every year!

Word Embeddings: Word2Vec(tor)

Word2Vec (Mikolov *et al.* 2013)

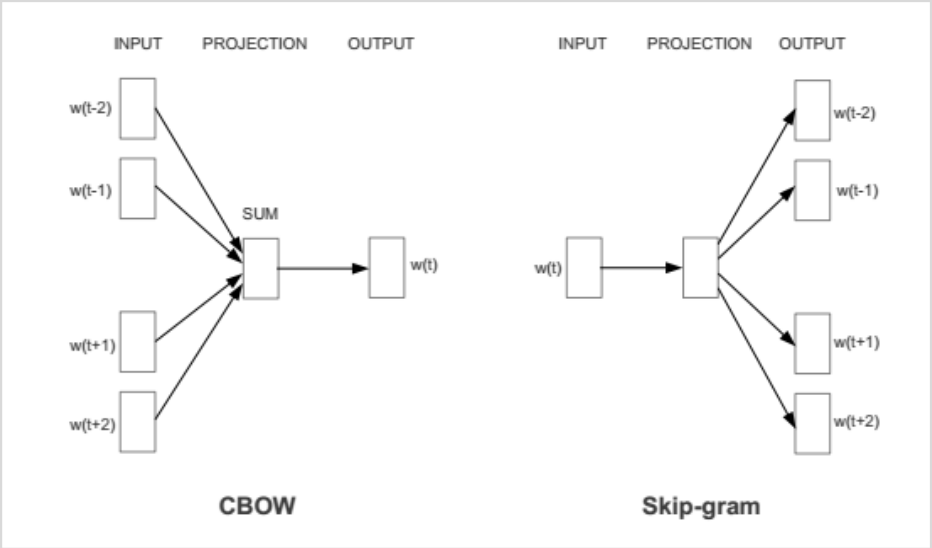
- For learning word representations using a shallow neural network
- No shared parameters across the network
- Fast training
  - Training over Wikipedia snapshot (about 1 billion words) with 50 dimensions takes only 4 hours
- Strong representations – previous state-of-the-art performance.
- Learned word embedding vectors can be represented as linear translations etc.

$$\text{vec(king)} - \text{vec(man)} + \text{vec(woman)} \cong \text{vec(queen)}$$
$$\text{vec(Paris)} - \text{vec(France)} + \text{vec(China)} \cong \text{Beijing)}$$

Word Embeddings: Word2Vec(tor) Continued.

Word2Vec Configurations

- Word2vec has two models
- Continuous Bag of Word (CBOW)
  - Predict a word *w(t)* given its context
- Skip-gram models
  - Predict context words given a word *w(t)*

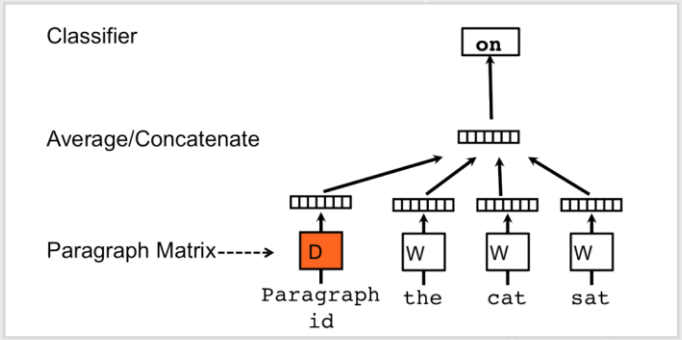


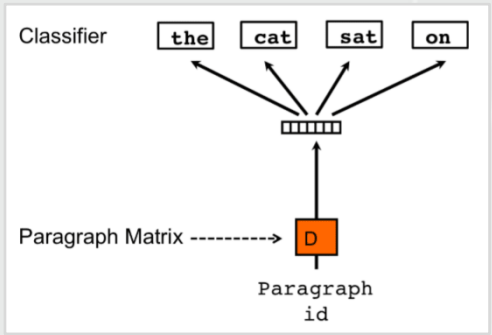
Other Language Representations

Different Embeddings Forms

- Sentences - Sent2Vec
- Paragraphs - Doc2Vec
- Document topics - Top2Vec
- Multi-Modal - Data2Vec

Document embedding





Agenda (Hands-on Session) – Word Representations

Learning word representations from scratch

- General domain and domain-specific word embeddings
- Word similarity
- Word vector clustering and visualisation

Supplementary Content

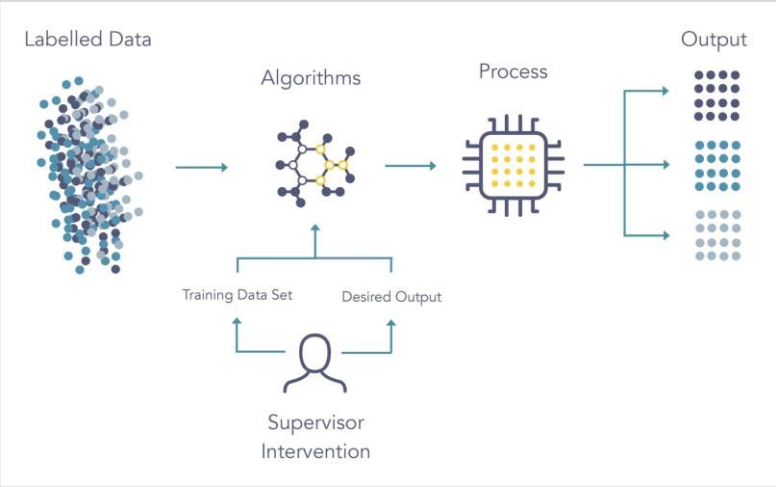
- Interactive exploration of word2vec models: https://ronxin.github.io/wevi/


12.3 Supervised Learning
Learning from example

Supervised Learning

Fundamentals of Supervised Learning

- **Aim**: Learn a complex function that maps X to y when provided a set of labelled x, y pairs as example
- x,y pairs are typically acquired through human annotation using a pre-defined schema/model
- Can vary in difficulty and complexity



Supervised Learning

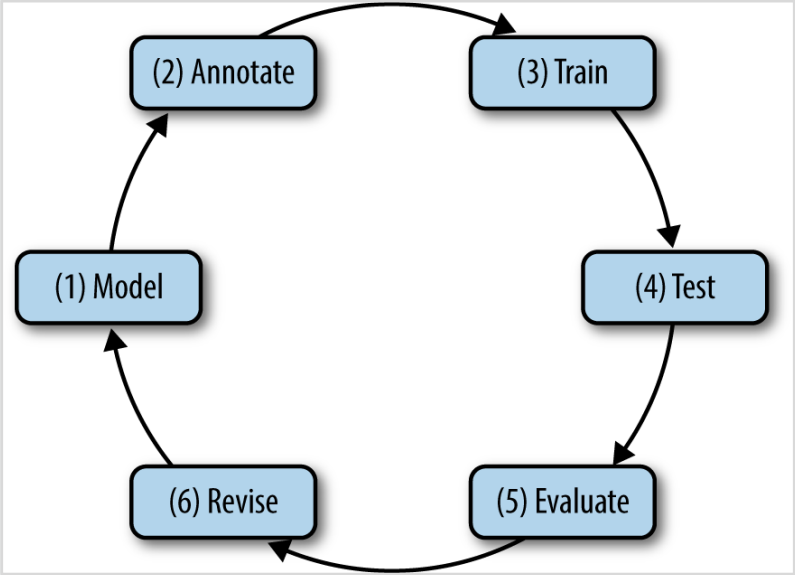
Agenda (Hands-on Session) – Exploring Supervised NLP Tasks using HuggingFace 🤗

- Sequence Classification
- Question Answering
- Language Modelling
- Text Generation
- Named Entity Recognition
- Summarization
- Translation
- Colab Notebook

Associate Prof. Wei Liu & Tyler Bikaun, UWA

Fundamentals of Natural Language Annotation

The Annotation Development Cycle



Annotation Example: Accident Report (NER)

1 - Specifying a Model for Annotation

- What are the goals of the annotation task?
- What are the characteristics of the dataset? Is it representative of the goals of the task? Is it balanced? What is the quality?
- Does the task require subject matter experts or linguists?
- Does a complete or partial model already exist?

walking in car park from office to workshop and rolled ankle

injury | activity | body part

Initial Model

Associate Prof. Wei Liu & Tyler Bikaun, UWA

Annotation Example: Accident Report (NER)

injury activity body part

2 - Performing human annotation

- What tool will be used to perform annotation?
- How many annotators are required to arrive at a *gold standard*?
- What level of resources are committed to the annotation task?
- How much data will be sufficient for training and testing?

activity

walking in car park from office to workshop and rolled ankle

injury body part

Associate Prof. Wei Liu & Tyler Bikaun, UWA

Annotation Example: Accident Report (NER)

injury | activity | body part

3 & 4 - Training and Testing

- What model architecture is best suited for the chosen task?
- How will the texts be represented?
- How will the model architecture be optimised?
- What is the acceptable level of performance and expected bayes error rate?

activity

walking in car park from office to workshop and rolled ankle

injury body part

Associate Prof. Wei Liu & Tyler Bikaun, UWA

Annotation Example: Accident Report (NER)

injury | activity | body part

5 – Evaluation of Model Performance

- Does the model meet the acceptable level of performance?
- What does the model do well at, what does the model struggle with?

activity

walking in car park from office to workshop and rolled ankle

injury body part

Associate Prof. Wei Liu & Tyler Bikaun, UWA

Annotation Example: Accident Report (NER)

6 – Revising the Model

- Does the model meet the goals of the task? If not, what modifications need to be made?
- How are the annotators performing? Is the task too challenging?

walking in car park from office to workshop and rolled ankle

injury | activity | body part | location

New Model

Annotation Example: Accident Report (NER)

activity location location location

walking in car park from office to workshop

and rolled ankle

injury body part

Associate Prof. Wei Liu & Tyler Bikaun, UWA

Annotation Example: Accident Report (ET)

location
location/car park

location
location/office

location
location/workshop

walking in car park from office to workshop

activity
activity/walking

and rolled ankle

injury
injury/sprain

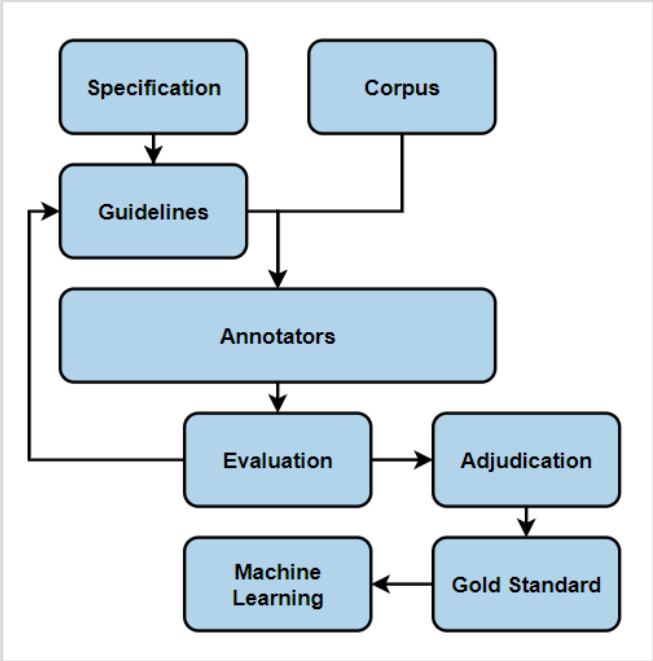
body part
body part/leg
body part/leg/ankle

Associate Prof. Wei Liu & Tyler Bikaun, UWA

Fundamentals of Natural Language Annotation

After Development - The Annotation Cycle

- **Specification**: model of phenomena for a specific task
- **Guidelines**: helps annotators reliably label or tag the corpus
- **Annotators**: recruited or crowd sourced humans
- **Evaluation**: inter-annotator agreement
- **Adjudication**: creating a gold-standard annotated corpus
- **Machine Learning**: using annotations for supervised learning




12.4 Unsupervised Learning
Learning without example

Unsupervised Learning: Topic Modelling

Suppose you have the following set of sentences:

- I **eat fish** and **vegetables**.
- *Fish* are *pets*.
- My *kitten* **eats fish**.

Latent Dirichlet allocation (LDA) is an unsupervised topic modelling algorithm that automatically discovers topics that these documents contain.

- **bold** words under the **Topic F**, which we might label as "**food**"
- *italics* words might be classified under a separate *Topic P*, which we might label as "*pets*"

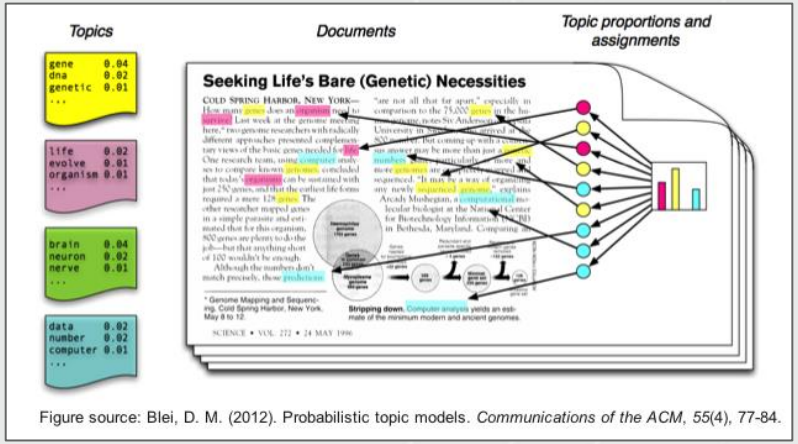


Figure source: Blei, D. M. (2012). Probabilistic topic models. *Communications of the ACM*, *55*(4), 77-84.

https://www.quora.com/What-is-the-best-way-to-explain-topic-modeling-to-a-layman

Unsupervised Learning: Topic Modelling

LDA defines each topic as a bag-of-words (bow), and you have to label the topics as you deem fit.

1. We can infer the content spread of each sentence by a word count
   - **Sentence 1** (I **eat fish** and **vegetables**): 100% Topic F
   - **Sentence 2** (*Fish* are *pets*): 100% Topic P
   - **Sentence 3** (My *kitten* **eats fish**): 33% Topic P and 67% Topic F
2. We can derive the proportions that each word constitutes in given topics. For example, Topic F might comprise words in the following proportions: 40% eat, 40% fish, 20% vegetables, …

Unsupervised Learning: Topic Modelling

Three Steps

LDA defines each topic as a bag-of-words (bow), and you have to label the topics as you deem fit.

1. Specify how many latent topics are there
2. For each word, assign a temporary topic, according to a Dirichlet distribution, e.g. kitten -> F
    - If a word appears twice, each word may be assigned to different topics.
    - Function words (e.g. “the”, “and”, “my”) are removed and not assigned to any topics.
3. Topic assignment is updated based on two criteria:
    - How prevalent is that word across topics?
    - How prevalent are topics in the document?

Unsupervised Learning: Topic Modelling

Repeated refinement

- The process of checking topic assignment is repeated for each word in every document, cycling through the entire collection of documents multiple times.
- This iterative updating is the key feature of LDA that generates a final solution with coherent topics.

	Document X		Document Y
	Fish		Fish
	Fish		Fish
	Eat		Milk
	Eat		Kitten
	Vegetables		Kitten

	Document X		Document Y
F	Fish	?	Fish
F	Fish	F	Fish
F	Eat	F	Milk
F	Eat	P	Kitten
F	Vegetables	P	Kitten

	Document X		Document Y
F	Fish	?	Fish
F	Fish	F	Fish
F	Eat	F	Milk
F	Eat	P	Kitten
F	Vegetables	P	Kitten

	Document X		Document Y
F	Fish	?	Fish
F	Fish	F	Fish
F	Eat	F	Milk
F	Eat	P	Kitten
F	Vegetables	P	Kitten

Unsupervised Learning

Agenda (Hands-on Session) – Topic Modelling using Latent Dirichlet Allocation (LDA)

- Corpus pre-processing including lemmatization
- Training LDA model from scratch
- Interactive visualisation of LDA model
- Exploration of LDA features

CORE Skills

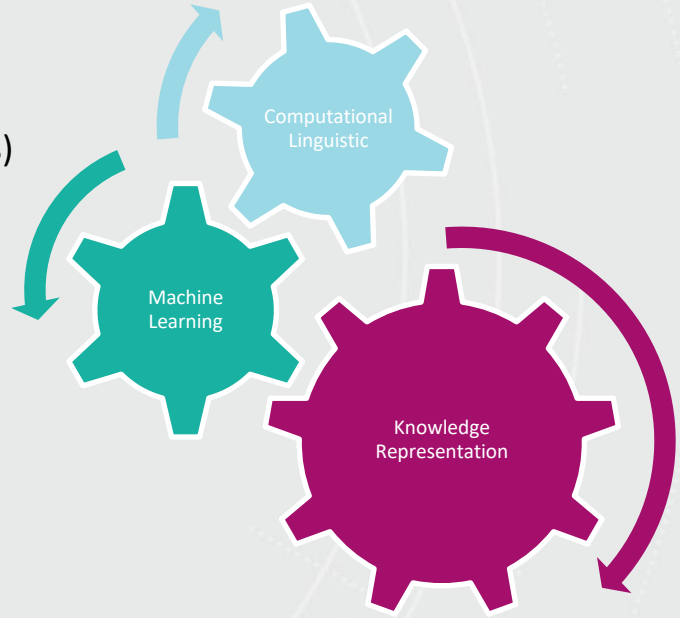
A CORE partnership

COREHUB.COM.AU/SKILLS

Additional Content

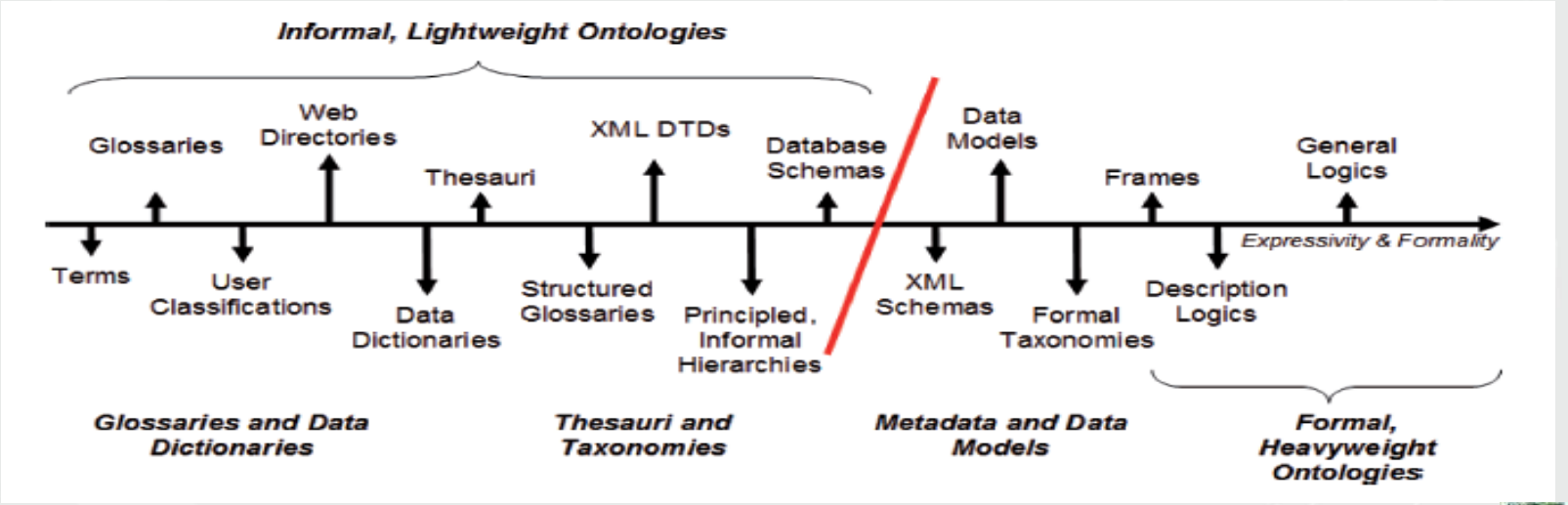
NLP Pipelines – Ontology Learning from Text

- **Graphs in knowledge representation**
  - Two different sources
    - Knowledge experts (logic based)
    - Knowledge discovery (from natural language texts)
- Ontology learning
  - Entity Extraction
    - Symbolic
      - Unsupervised learning
      - Supervised learning
    - Subsymbolic
      - Deep learning
  - Entity Relation Extraction



Lightweight Ontologies

Overview – What's Light Weight Ontology



Wilson Wong, Wei Liu, and Mohammed Bennamoun. (2012) *Ontology Learning from Text: A look back and into the future*, ACM COMPUTING SURVEYS, 44, 4, Article 20, pp. 20:1-20:36

text > sets of corpora are prepared for ontology construction. Web Crawling is performed to gather web pages from general sources such as Reuters, Discovery and CNet to create the Contrastive Corpus. Readily available non-domain specific

Lightweight Ontologies

Text Preprocessing → texts → **Text Processing** → term candidates → **Term Recognition** → terms → **Relation Acquisition**

- parse HTML/XML files for texts.
- segment texts to identify sentences.
- **remove noise**.

- parse sentences for part-of-speech and grammatical information.
- **extract noun phrases** as term candidates and measure their collocational strength (i.e. **unithood**).

- determine the relevance of term candidates with respect to a domain (i.e. **termhood**) using text corpora.

- **acquire semantic relations** between terms.

...University of WA... universities of Perth...

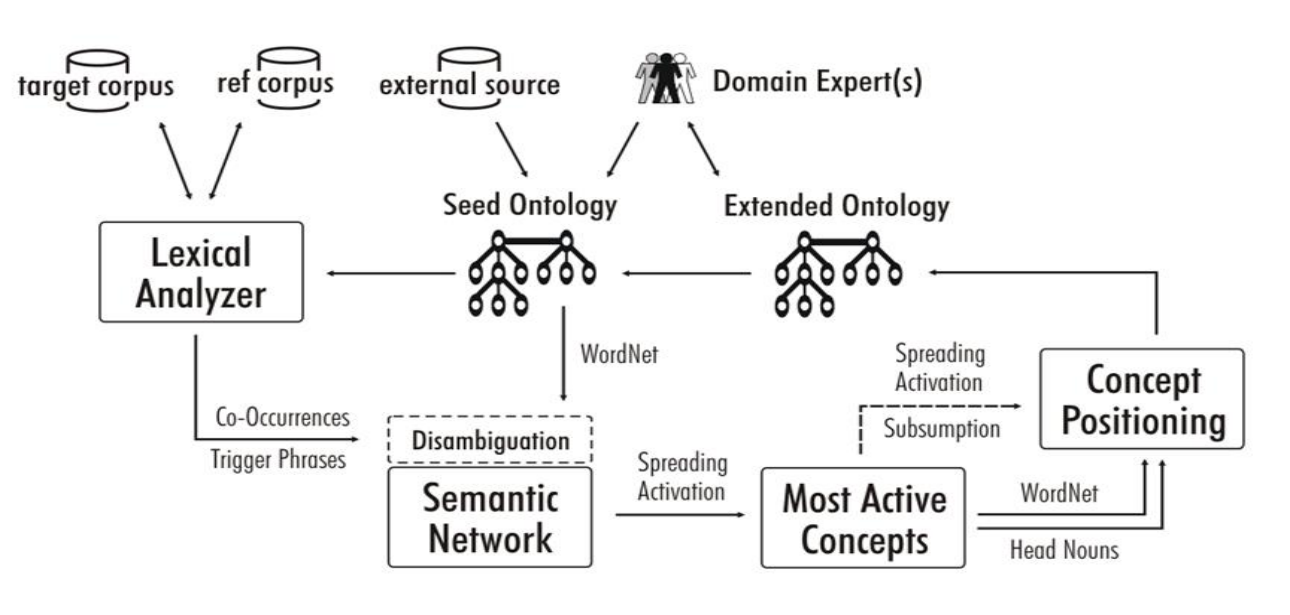
...[University of WA]...[universities] of [Perth]....

[University of WA]
[education]
[next year]
[Perth]
...

How is ***[University of WA]*** related to ***[education]***

Ontology Learning from Text – System Overview

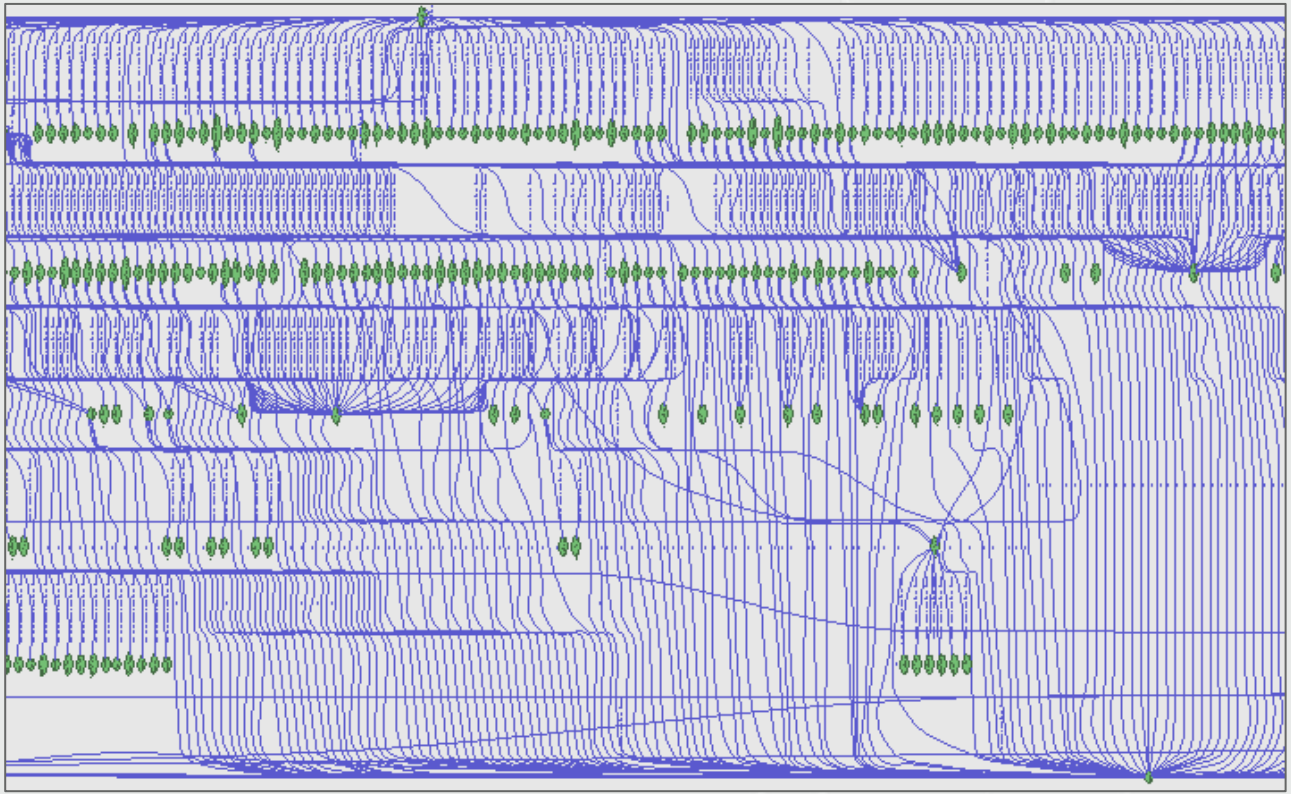
A semi-automatic ontology learning system



Wei Liu, Arno Scharl & Albert Weischselbraun, 2005

Ontology Learning from Text – Term Co-ocurrence

Semantic Network Visualisation



Ontology Learning from Text – Term Significance

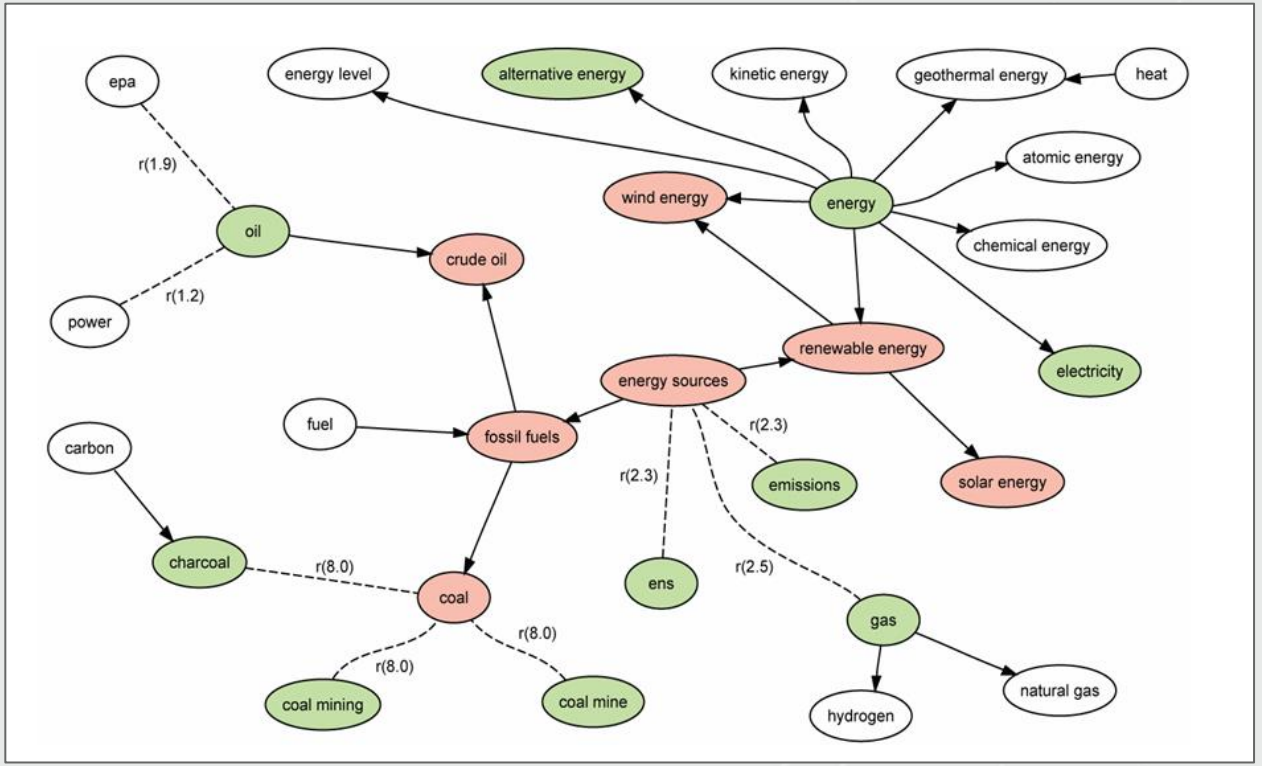
Results of climate change data analysis

ENERGY	
DOCUMENT LEVEL	SIG
gas	128,714
power	67,176
natural gas	62,396
natural	46,337
electricity	43,735
fuel	25,761
oil	25,168
allegheny	24,183
renewable	23,815
utility	23,554

ENERGY	
SENTENCE LEVEL	SIG
renewable energy	204,328
renewable	194,479
allegheny energy	185,018
allegheny	154,046
energy efficiency	147,639
duke energy	143,573
energys	137,306
sempra	119,689
sempra energy	105,757
centerpoint	101,068

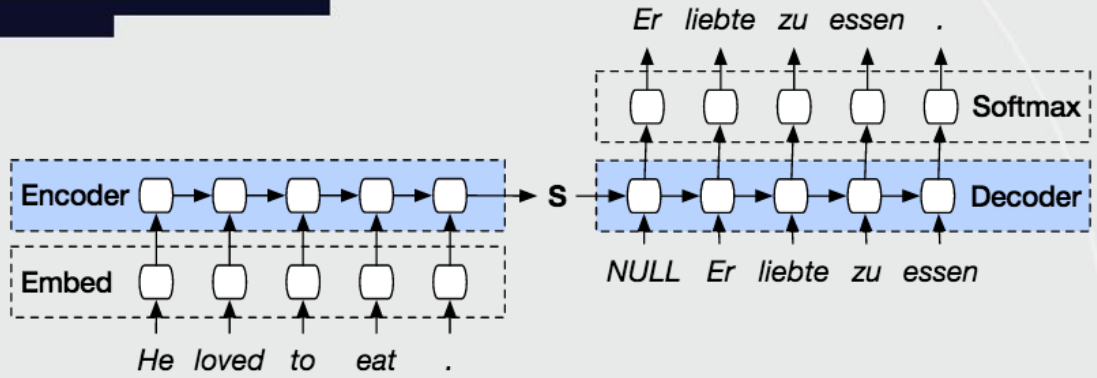
Ontology Learning from Text – Building Lightweight Ontologies

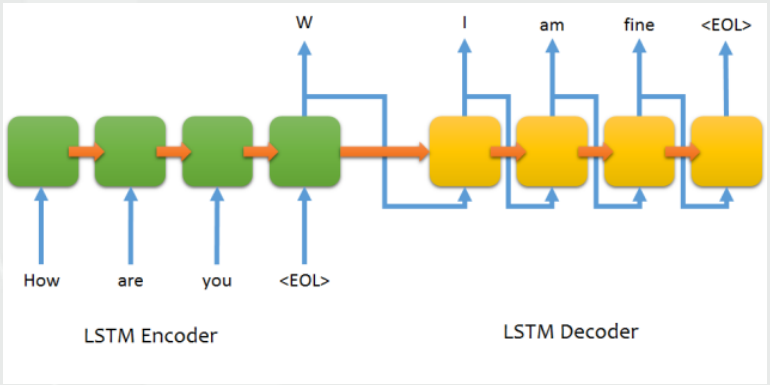
Spreading Activation

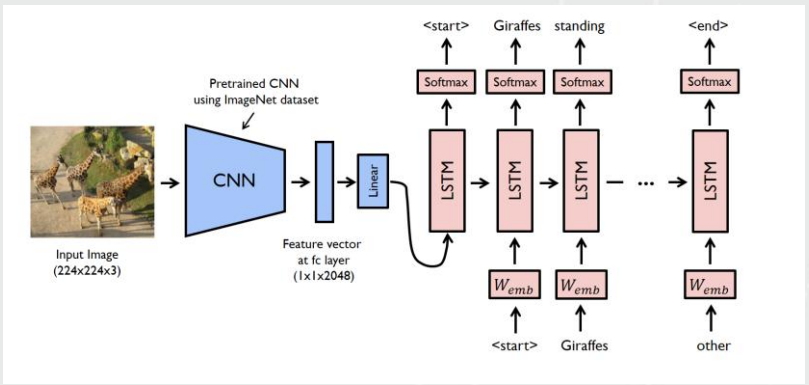


12.4 Neural Language Processing

Sequence to Sequence Learning (Encode then Decode)

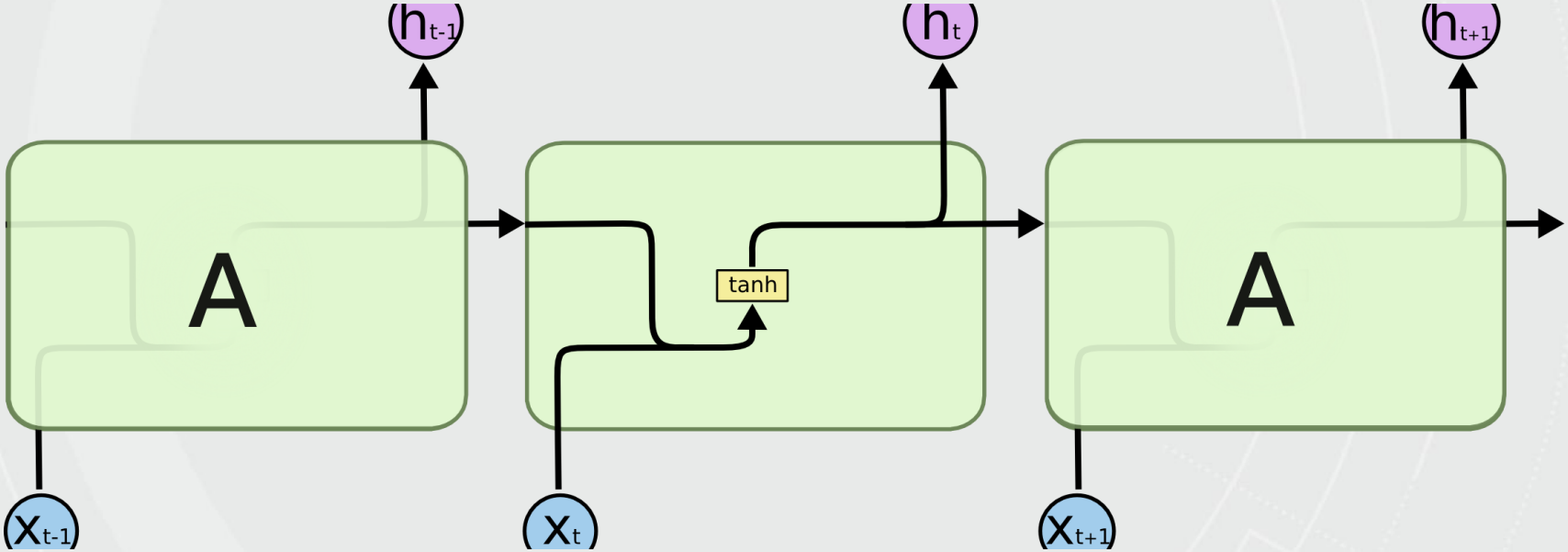






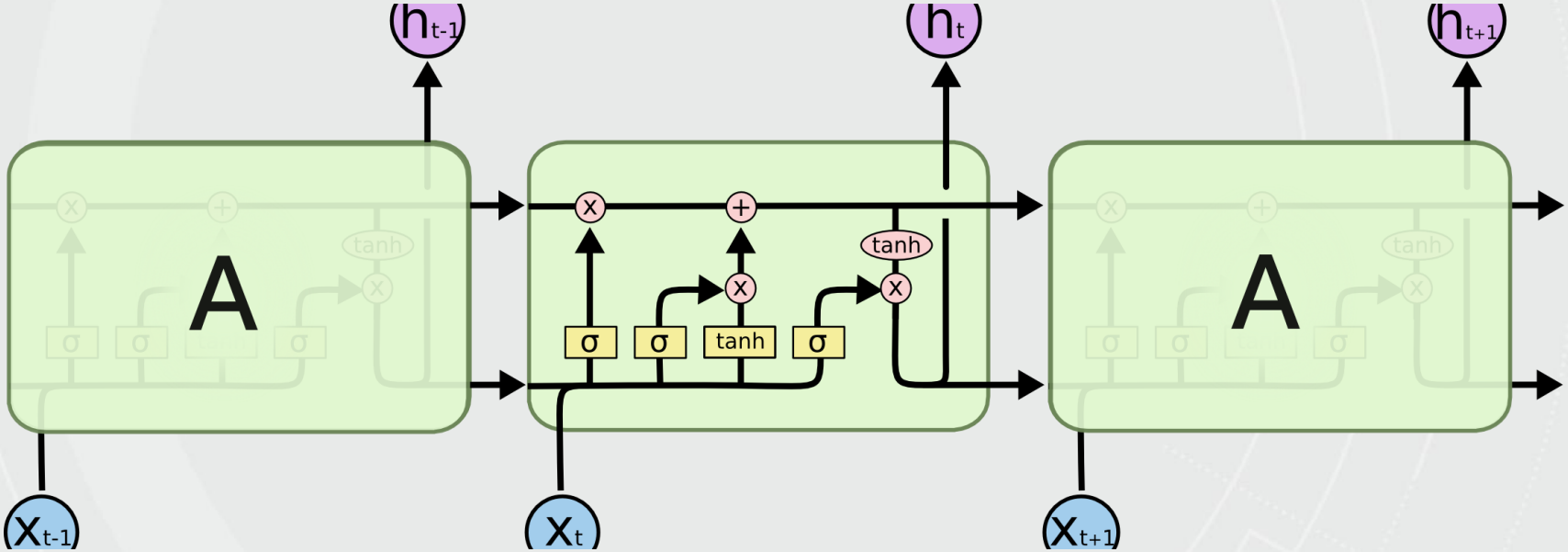
LSTM

Contains four interacting layers

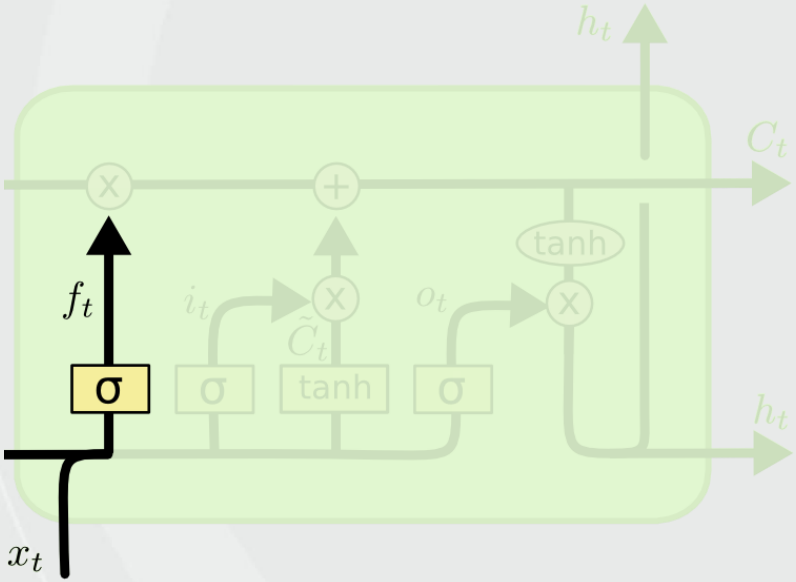


LSTM

Inside a cell



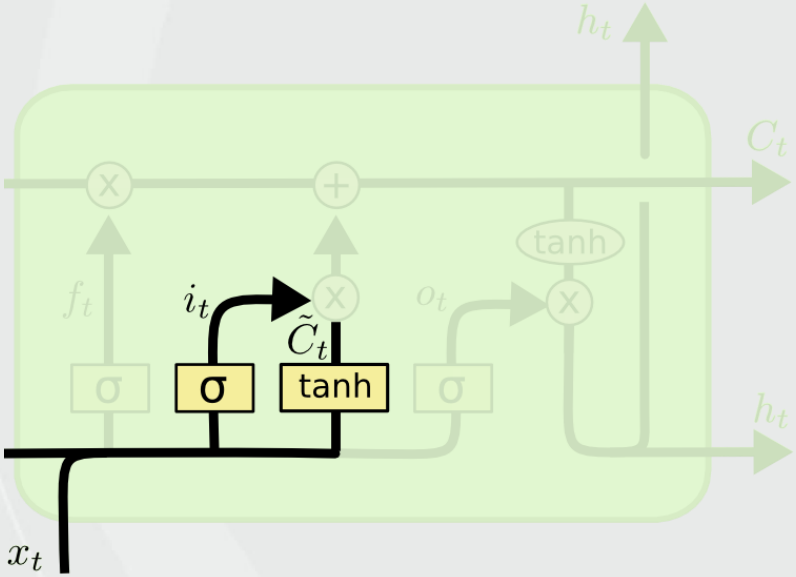
LSTM – Forget gate



$$f_t = \sigma\left(W_f \cdot [h_{t-1}, x_t] \; + \; b_f\right)$$

http://colah.github.io/posts/2015-08-Understanding-LSTMs/

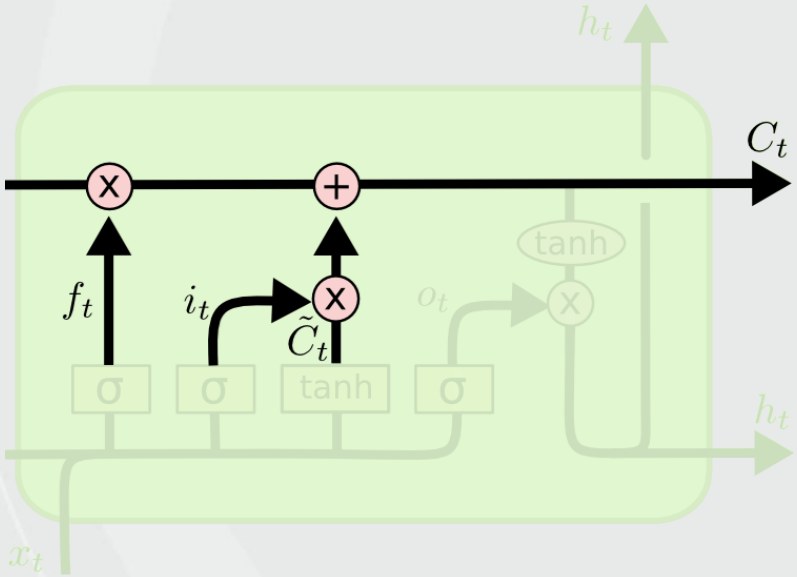
LSTM – Input Gate Layer



$$i_t = \sigma\left(W_i \cdot [h_{t-1}, x_t] \; + \; b_i\right)$$

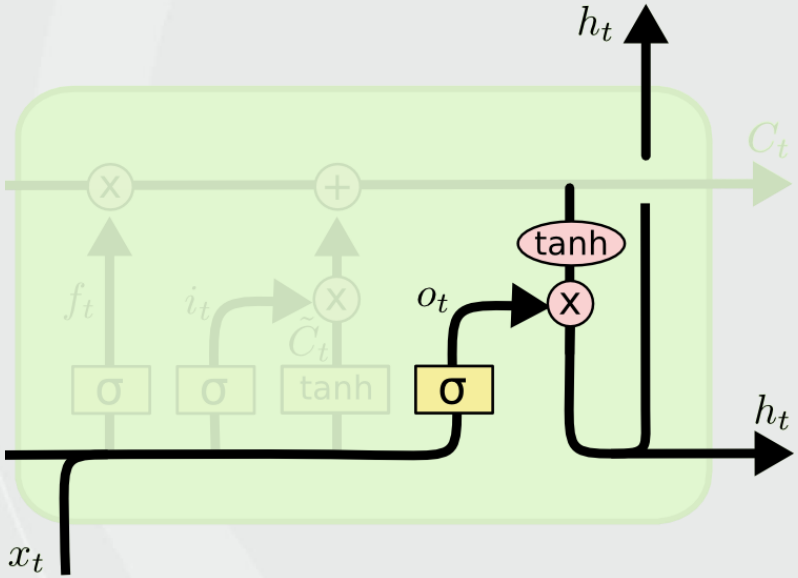
$$\tilde{C}_t = \tanh(W_C \cdot [h_{t-1}, x_t] \; + b_C)$$

LSTM - Update



$$C_t = f_t * C_{t-1} + i_t * \tilde{C}_t$$

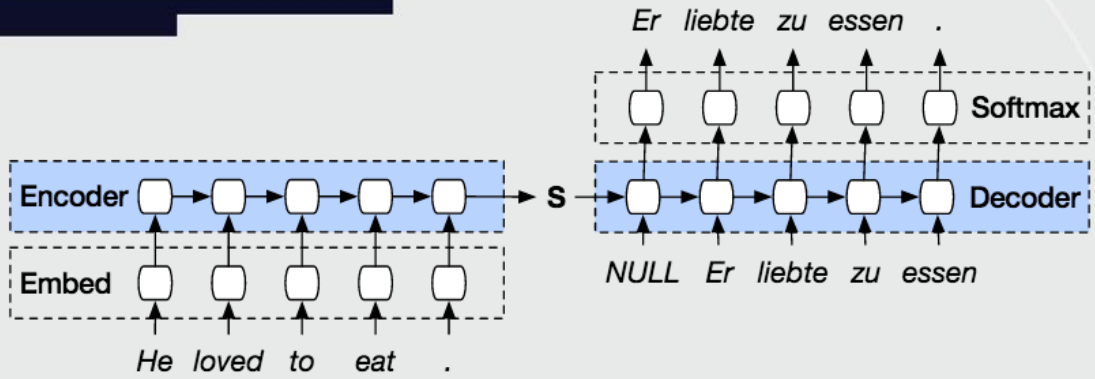
LSTM – Output Gate

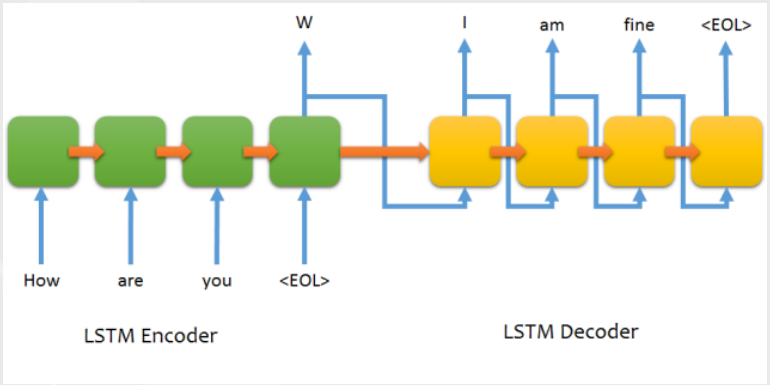


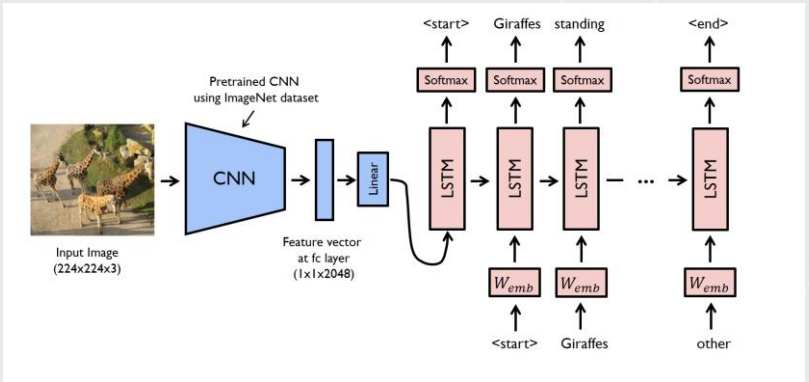
$$o_t = \sigma\left(W_o\left[h_{t-1}, x_t\right] + b_o\right)$$

$$h_t = o_t * \tanh\left(C_t\right)$$

Sequence to Sequence Learning (Encode then Decode)







Associate Prof. Wei Liu

Educator

Program Chair of Data Science, Department of Computer Science and Software Engineering, The University of Western Australia

CORE Skills Data Science Springboard Delivery Team

Wei Liu is a teaching and research academic at UWA in the Department of Computer Science and Software Engineering focusing on knowledge discovery from natural language text. Her research focuses on knowledge discovery from natural language text, deep learning methods for knowledge graph construction and analysis, as well as sequential data mining and forecasting in traffic and water consumption domain.

Wei received her PhD from the University of Newcastle, Australia in 2003. She has published in highly reputable journals such as ACM Computer Surveys, Journal of Data Mining and Knowledge Discovery, Knowledge and Information Systems and presented at key international events including the International Conference on Data Engineering (ICDE) and ACM International Conference on Information and Knowledge Management (CIKM).

Wei has won three Australian Research Council Grants and managed several industry grants. Her current industry-related research projects include knowledge graph refinement for geological survey reports, incident log analysis and visualisation, short-term traffic predication and cognitive computing for asset management.

Program Delivery Day 12 (Special Data Types - Natural Language Processing & Text Mining)

Email: wei.liu@uwa.edu.au | Publications: https://research-repository.uwa.edu.au/en/persons/wei-liu

Tyler Bikaun

Educator

PhD Candidate (Technical Language Processing), Department of Computer Science and Software Engineering, The University of Western Australia & Mechanical Engineer

CORE Skills Data Science Springboard Delivery Team

Tyler Bikaun is a Mechanical Engineer turned PhD candidate at UWA and a scholar with the Mineral Research Institute of Western Australia and ARC Training Centre for Transforming Maintenance through Data Science. His research focuses on the intersection of computer science and industrial engineering, in particular, applying computational techniques to technical texts in the context of industrial maintenance.

Prior to commencing his PhD, Tyler worked in engineering consulting to the resources and utilities sectors in Western Australia. Here he was involved in asset performance optimisation of rotating equipment and advanced analytics for asset health monitoring applications.

Program Delivery Day 12 (Special Data Types - Natural Language Processing & Text Mining)

Email: tyler.bikaun@research.uwa.edu.au | LinkedIn: https://www.linkedin.com/in/tyler-bikaun