CORE Skills

Delivering Data Science
In Resources & Energy

Knowledge Discovery from Natural Language Text

Day 12: Special Data Tyles – Natural Language Processing and Text Mining

Finding Needles in Wordstacks

A/Prof. Wei Liu & Tyler Bikaun

Computer Science and Software Engineering

wei.liu@uwa.edu.au
tyler.bikaun@research.uwa.edu.au

THE UNIVERSITY OF WESTERN AUSTRALIA

A CORE partnership

Program Timeline

Day 12: Special Data Types - Natural Language Processing & Text Mining

Leader Engagement	15 Day Professional Program														
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 - Natural Language Processing and Text Mining		Special Data	Capstone Project Development & Presentation	Capstone Propeller
2-hours	Day 1	Day 2	Day 3	Day 4	Day 5	Day 6	Day 7	Day 8	Day 9	Day 10	Day 11	Day 12		Day 13	Day 14	Day 15
January 2022	Wed 24 Nov 2021	Thurs 25 Nov 2021	Wed 1 Dec 2021	Wed 8 Dec 2021	Wed 15 Dec 2021	Wed 19 Jan 2022	Wed 2 Feb 2022	Wed 9 Feb 2022	Wed 16 Feb 2022	Wed 23 Feb 2022	Wed 2 Mar 2022	Wed 9 Mar 2022		Wed 16 Mar 2022	Wed 23 Mar 2022	Wed 22 June 2022
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

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

Program Timeline

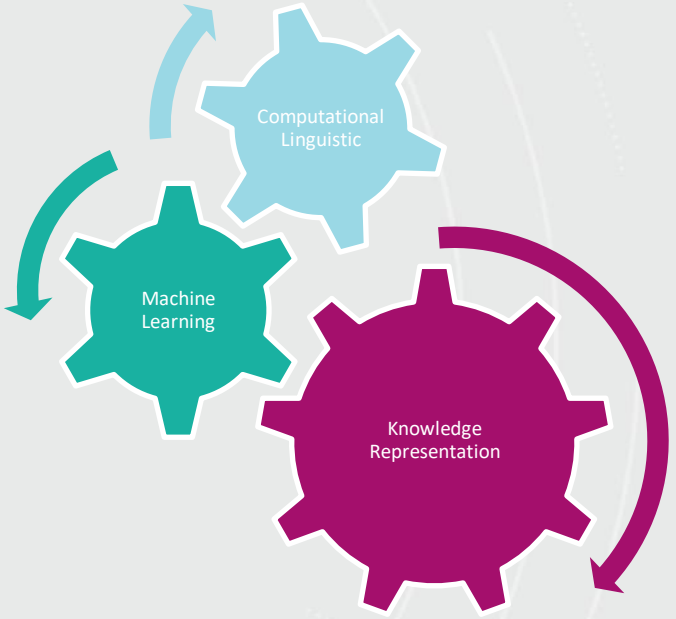
Day 12: Special Data Types - Natural Language Processing & Text Mining

Prerequisite		Introduction to Data Projects	Data Analysis			Data & Communication Sandbox	Data Fusion and Machine Learning		Data Fusion Sandbox	Special Data Types - Time - series Data	Special Data Types - Natural Language Processing and Text Mining	Special Data Types - Spatial Data	Capstone Project Development & Presentation	Capstone Propeller
Day 1	Day 2	Day 3	Day 4	Day 5	Day 6	Day 7	Day 8	Day 9	Day 10	Day 11	Day 12	Day 13	Day 14	Day 15
Thurs 30 Jan 2020	Fri 31 Jan 2020	Thurs 6 Feb 2020	Thurs 13 Feb 2020	Thurs 20 Feb 2020	Thurs 27 Feb 2020	Fri 28 Feb 2020	Thurs 5 Mar 2020	Thurs 12 Mar 2020	Fri 13 Mar 2020	Thurs 19 Mar 2020	Thurs 26 Mar 2020	Thurs 2 Apr 2020	Fri 3 Apr 2020	Thurs 2 July 2020
Introduction to the program tools - Where Data Science comes from	Introduction to the program tools - The Why	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	Sandbox - Consolidate approaches covered and test on datasets	The 4th dimension and predictions	Finding needles in wordstacks	Spatial analytics and predictions	Pitching Capstone Projects	Project Review Day

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

Schedule

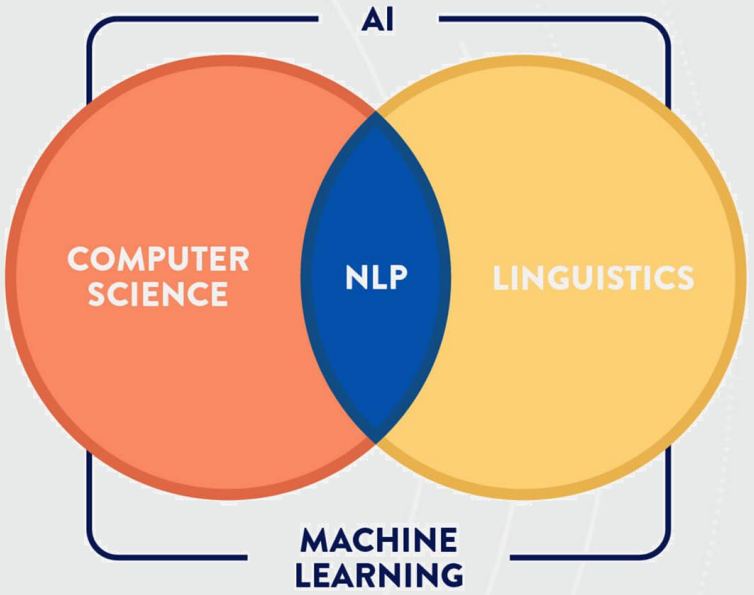
AWST	AEST	Agenda	Facilitator(s)
07:30	09:30	**Q&A, Issues & Announcements**	Tamryn/Wei/Tyler
07:45	09:45	12.0 Overview 12.1 Fundamentals of NLP	Wei Tyler
09:15	11:15	*Morning Tea*	
09:30	11:30	12.2 Fundamentals of NLP 12.3 Supervised Learning	Tyler
11:00	13:00	*Lunch*	
11:45	13:45	12.3 Supervised Learning	Tyler
13:15	15:15	*Afternoon Tea*	
13:30	15:30	**Project Update**	All/Tamryn
14:00	16:00	12.4 Unsupervised Learning	Tyler
14:30	16:30	**Closeout & Takeaways**	Wei/Tyler
14:55	16:55	**Menti Feedback**	Tamryn
15:00	17:00	**Close**	


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
- …

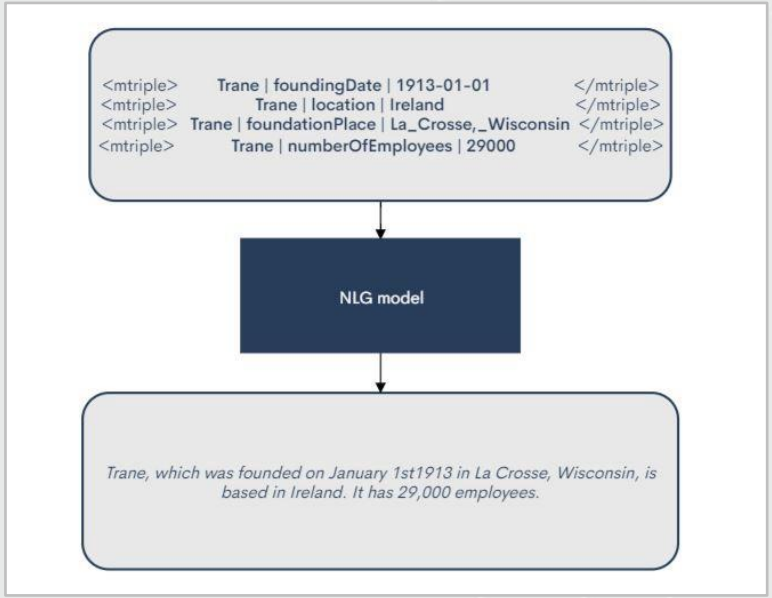


1 - http://nlpprogress.com/

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** !

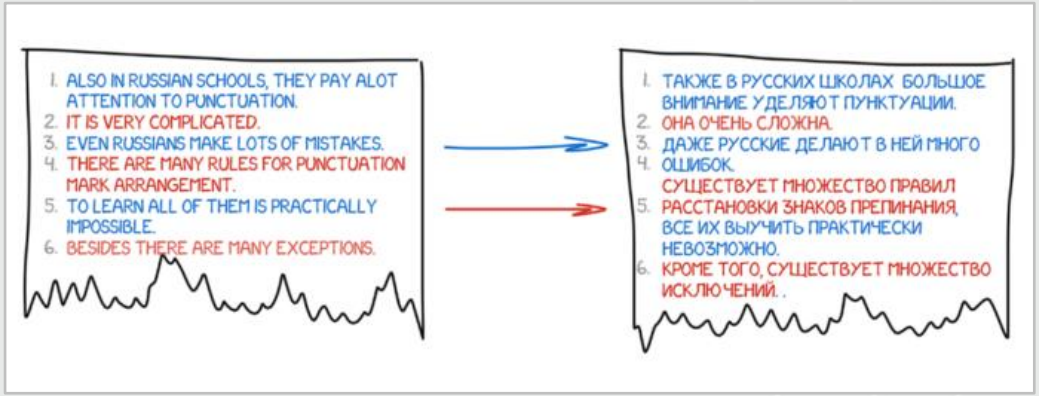
target: got **extra** to share, **are you** interested? Concert in **hong kong** !

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
- …

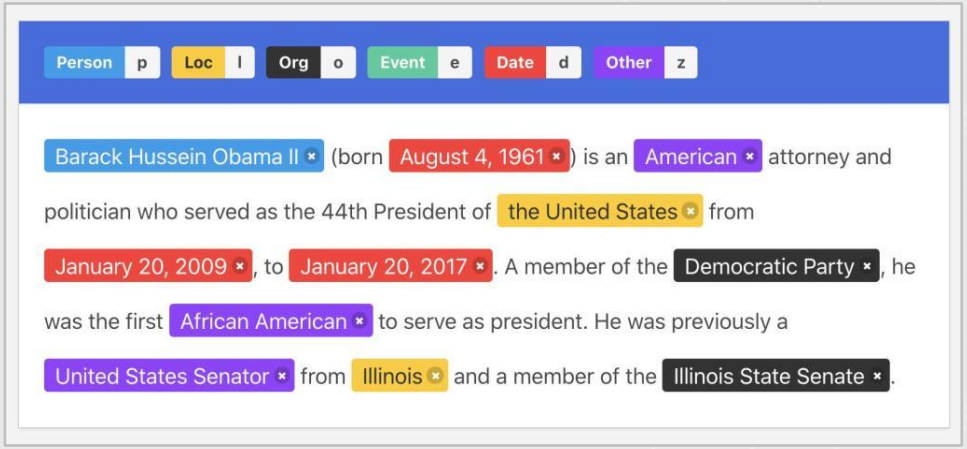


1 - http://nlpprogress.com/

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
- ...

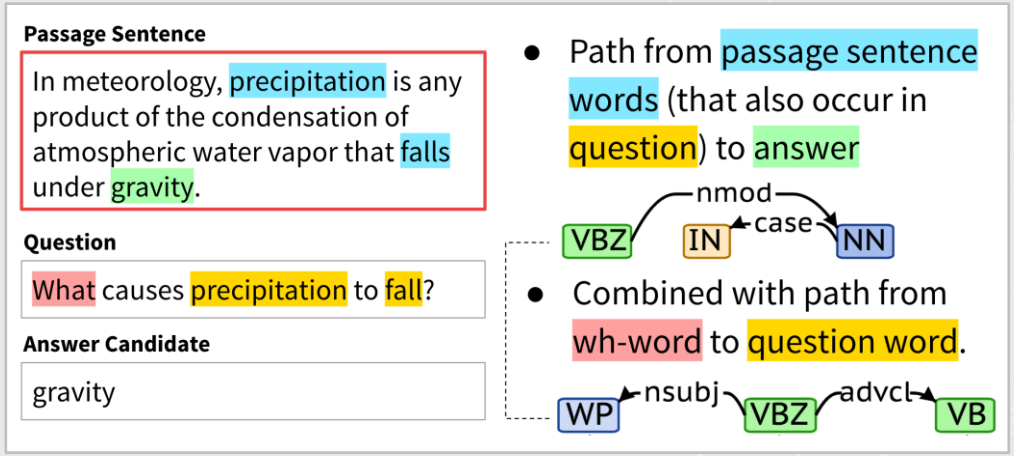


1 - http://nlpprogress.com/

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
- ...

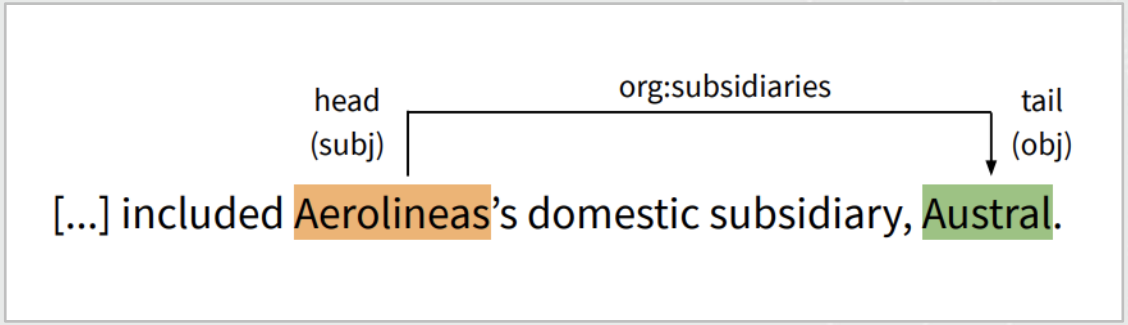


1 - http://nlpprogress.com/

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
- …

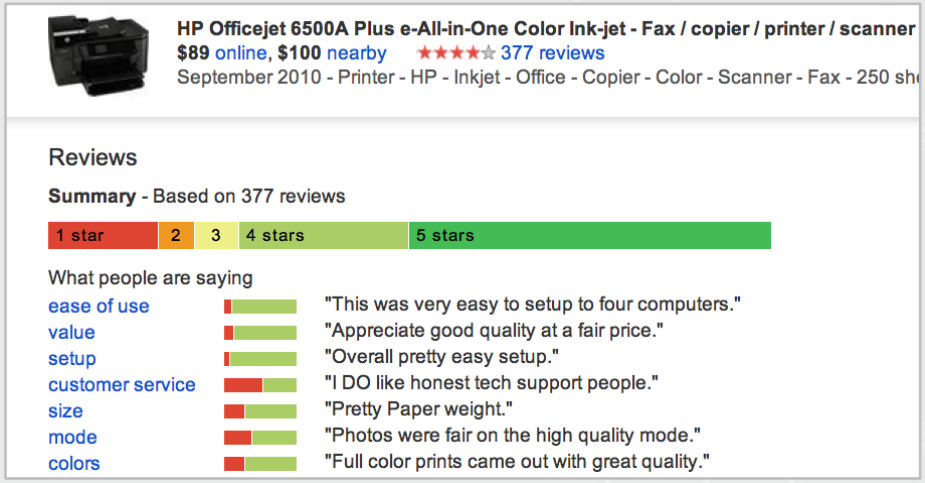


1 - http://nlpprogress.com/

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
- ...

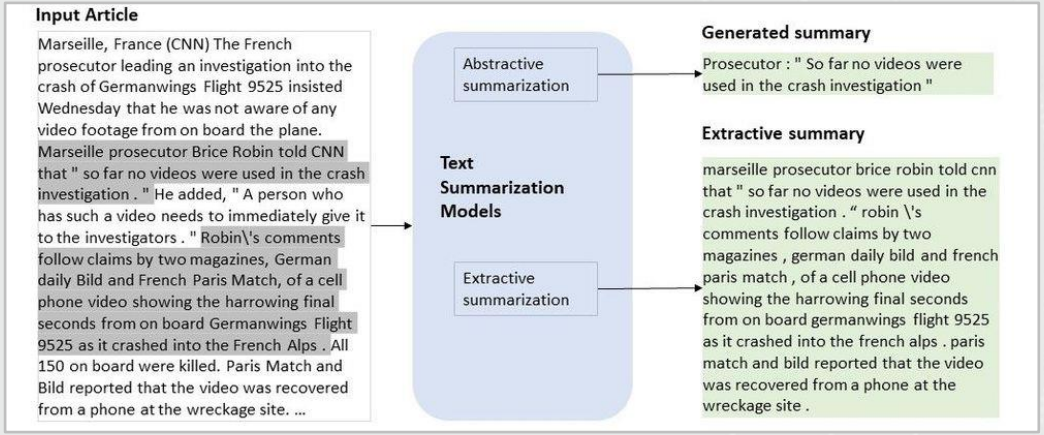
Example of a Complex Sentence	Example of a Simplified Sentence
Grammarly provides assistance in order to optimize users' communication.	Grammarly helps people communicate.

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
- ...

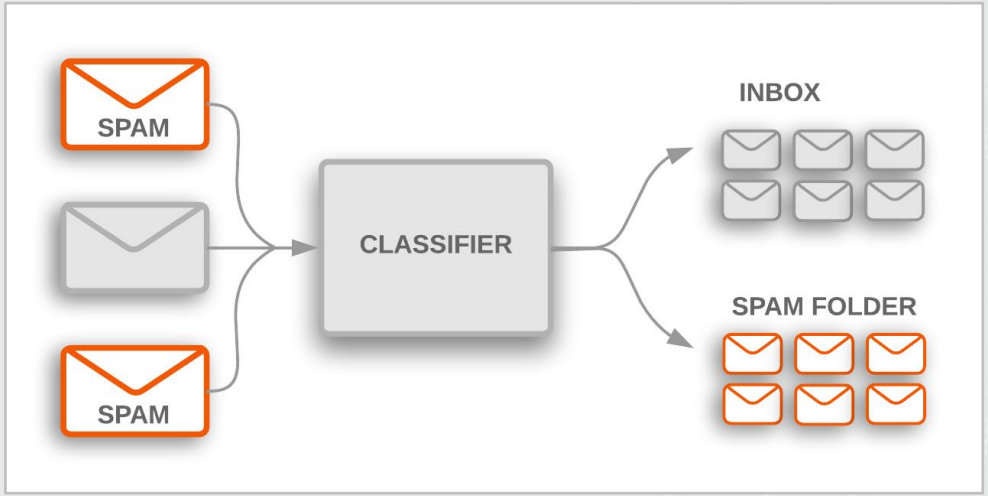


1 - http://nlpprogress.com/

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

Natural and Technical Language Processing

Natural Language Processing

- **Natural language processing** (NLP) is the study of getting computers to understand natural languages.
- You probably make use of NLP every day!
  - Google search
  - Spellcheckers
  - Siri
  - Autocomplete

Technical Language Processing

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

Accident reports:	WAlking in car park from office to workshop & roled ankle
	Injured right wrist lifting wheel barrow off ute
Maintenance work orders:	a/c blowing hot air
	repace grouser plate to l/h side track

Associate Prof. Wei Liu & Tyler Bikaun, UWA

NLP-TLP Research Group at UWA: http://nlp-tlp.org/

Natural and Technical Language Processing

Our Team

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

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

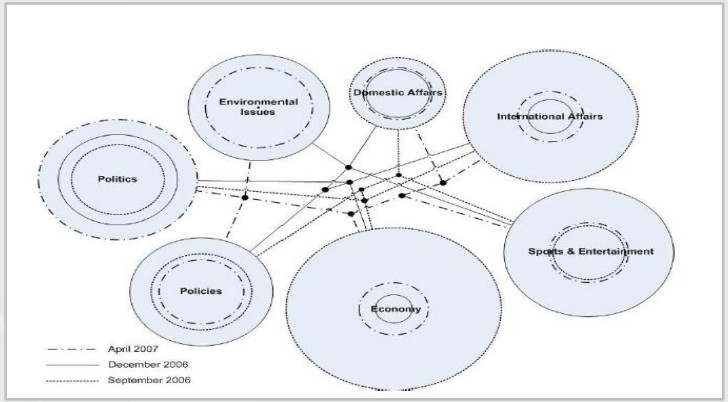
+ masters and honours students

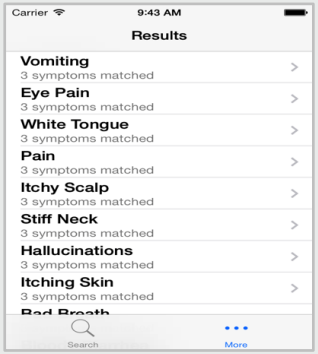
Associate Prof. Wei Liu & Tyler Bikaun, UWA

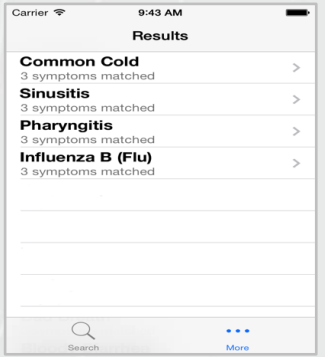
Past Projects

News, Academic Publication and Medical Diagnosis

- Election predication (news article analysis)
- Medical diagnosis
- PubMed Article Analysis







Associate Prof. Wei Liu & Tyler Bikaun, UWA

Past Projects

Geological Information Extraction

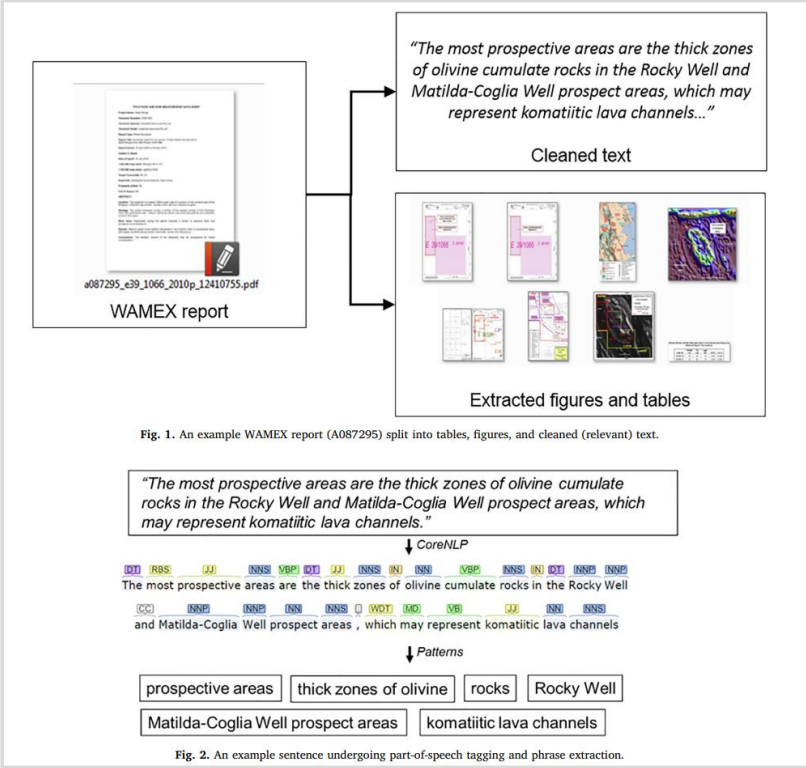


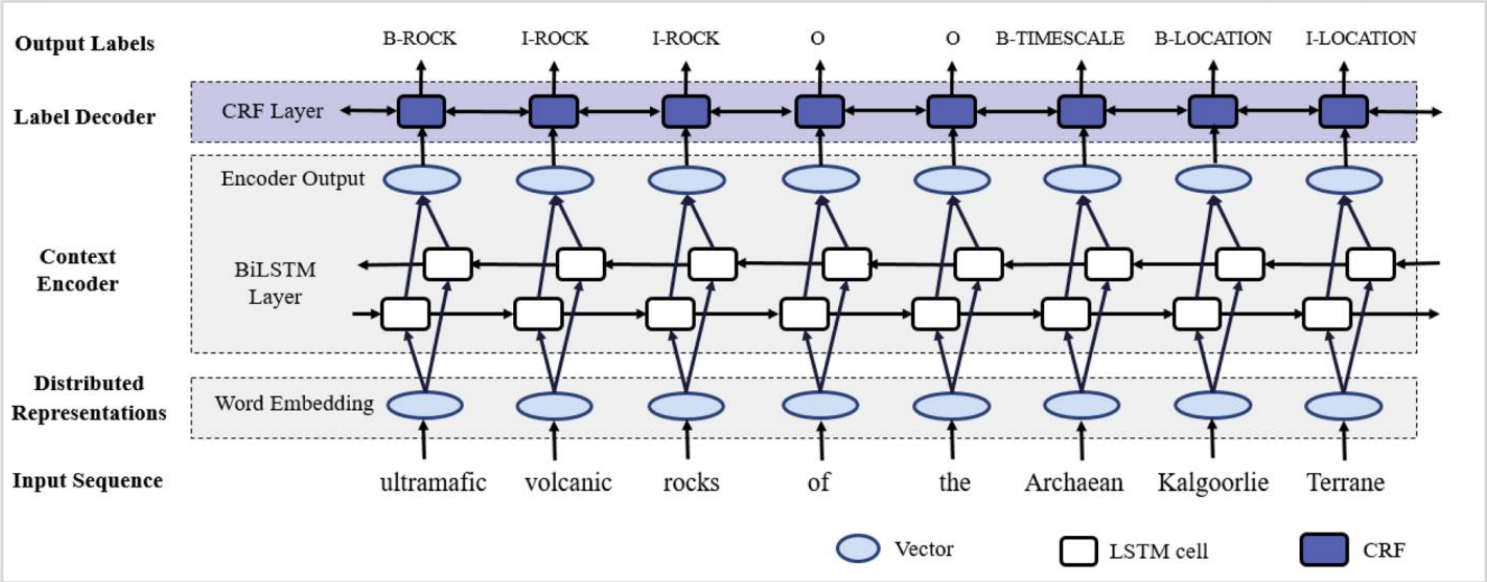
Fig. 1. An example WAMEX report (A087295) split into tables, figures, and cleaned (relevant) text.

Fig. 2. An example sentence undergoing part-of-speech tagging and phrase extraction.

Associate Prof. Wei Liu & Tyler Bikaun, UWA

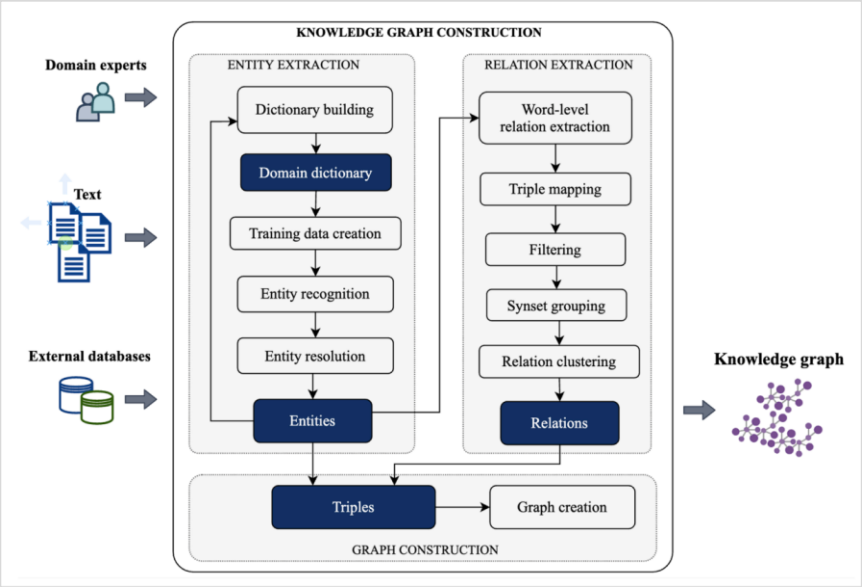
Past Projects

Geological Information Extraction



Past Projects

Geological Information Extraction



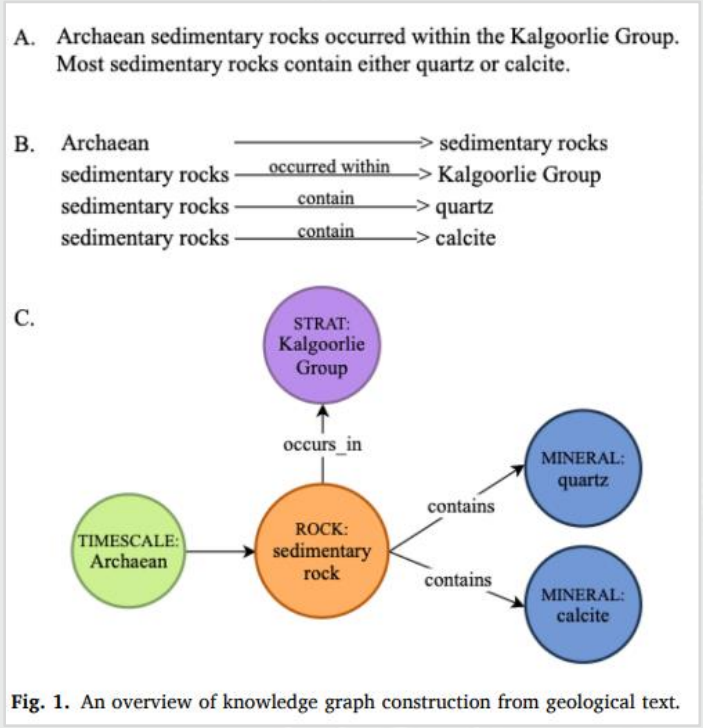


Fig. 1. An overview of knowledge graph construction from geological text.

Associate Prof. Wei Liu & Tyler Bikaun, UWA

Past Projects

Aquila – Knowledge Discovery from Mining Safety Data

- Aquila is a **web-based tool** for **visualising and analysing** technical records.
- It is designed to be **usable by anyone** – no programming experience required.
- It features a range of **unique visualisations** that support **knowledge discovery**.

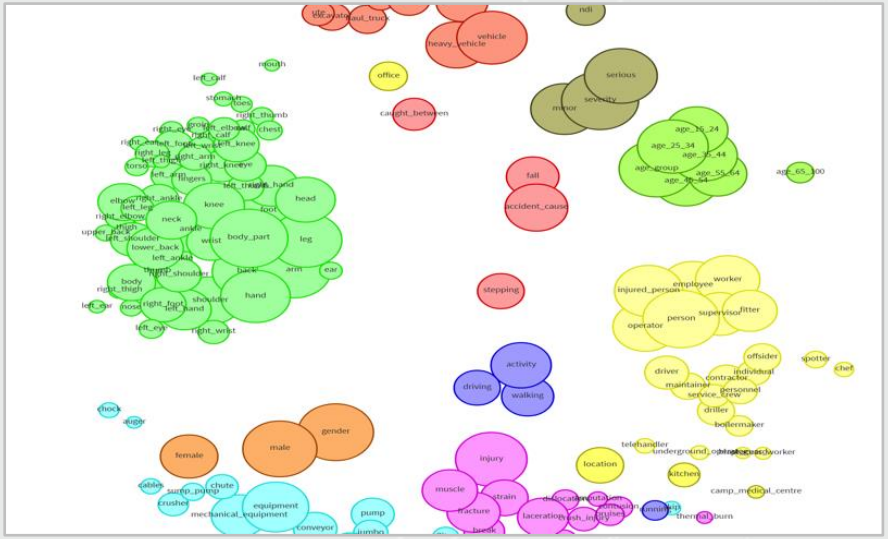
Associate Prof. Wei Liu & Tyler Bikaun, UWA

S³ - Semantics

Extracting Semantics – Association Rule Mining

- Aquilia – Text Mining (https://nlp-tlp.org/aquila)

AF	AG	
Contractor Company		Detailed Description
Contractor		Traumatic crush injury to the head and chest.
Contractor		Two fitters were undertaking maintenance work when an accident occurred resulting
Contractor		At approximately 9.20am on Wednesday 4th December a contractor working on the t
Contractor		Drill fitter had finished planned services for the shift and had returned the tools to th
Contractor 1		The injured person identified a stinging sensation on his left shin area. The injured pe
Contractor 1		Injured person was sleeping when he felt a stinging sensation on his right thigh area.
Contractor 10		An Operator was transporting signs in a loader bucket. Whilst removing one of the sig
Contractor 10		On 11/11/13 at approximately 0900hrs employee was assisting two other operators i
Contractor 11		Operator was climbing onto the rear of the bogger to straighten a hand rail when it g
Contractor 11		Service crew member stepped down from work cage onto uneven ground and rolled
Contractor 11		walking in car park from office to workshop & rolled ankle
Contractor 11		The bogger operator lost his footinng whilst climbing down off his loader, causing a t
Contractor 11		Person was walking around the bogger in the workshop and trod in some grease that



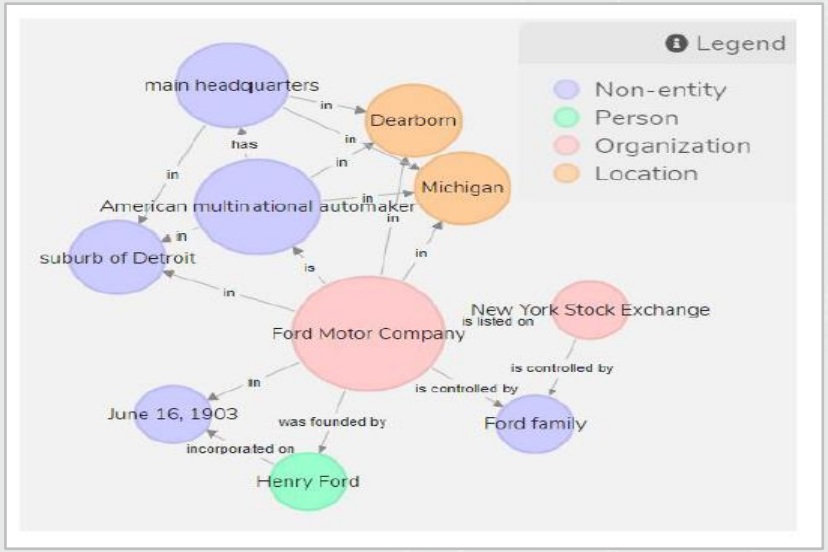
Associate Prof. Wei Liu & Tyler Bikaun, UWA

S³ - Semantics

Extracting Semantics – Knowledge Graph Construction from Text

- Text2KG (https://nlp-tlp.org/text2kg)

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.

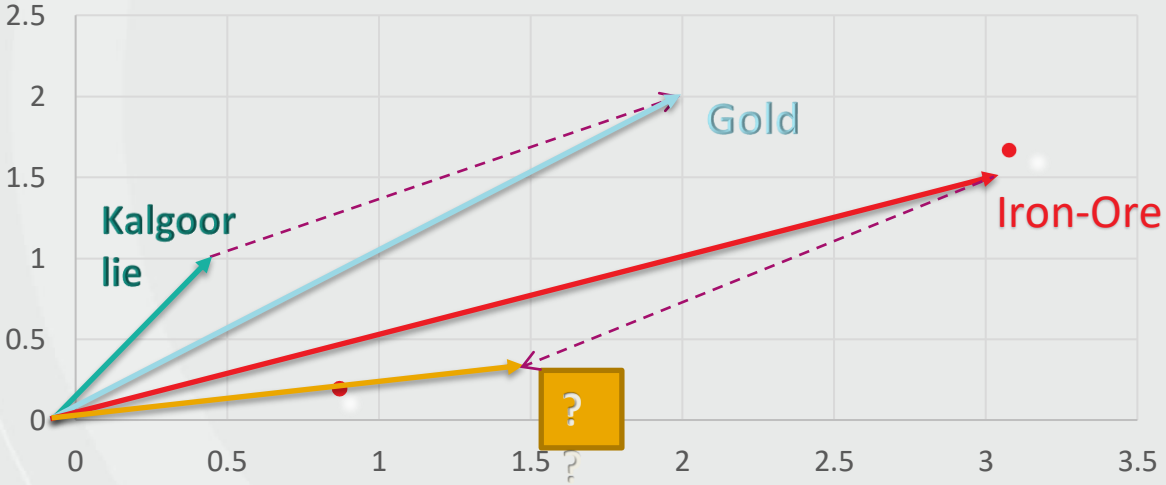


Associate Prof. Wei Liu & Tyler Bikaun, UWA

S^3 – Social

Extracting Social Relations between words

- Kalgoorlie + iron-ore – gold = ?



Associate Prof. Wei Liu & Tyler Bikaun, UWA

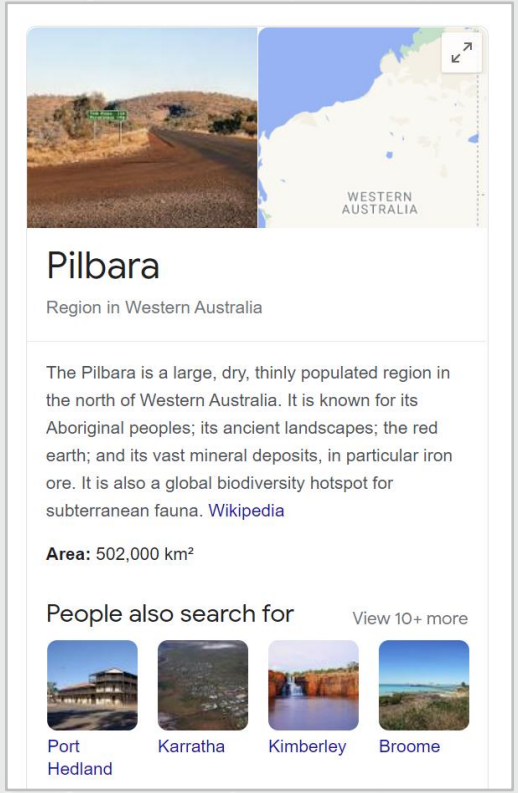
S^3 - Social

Extracting Social Relations between words

Relation `Commodity:Location`. Query: `kalgoorlie` + `iron-ore` − `gold`

Embedding	Top answers
Word2Vec raw	*pannawonica, windarling, mmif, ravensthorpe, karratha, newman*
FastText raw	*esperance, geraldton, bunbury, hyden, karratha*
Word2Vec terms	*martite, marandoo, iron, west_angelas, hematite, windarling*
FastText terms	*hematite, west-angelas, windarling, cunderdin, marandoo, iron*
Word2Vec pre-trained	*Pilbara, Pilbara_region, Port_Hedland, Pilbara_iron_ore, Karratha*
FastText pre-trained	*Pilbara, Karratha, Oakajee, Middlemount, Nullagine*

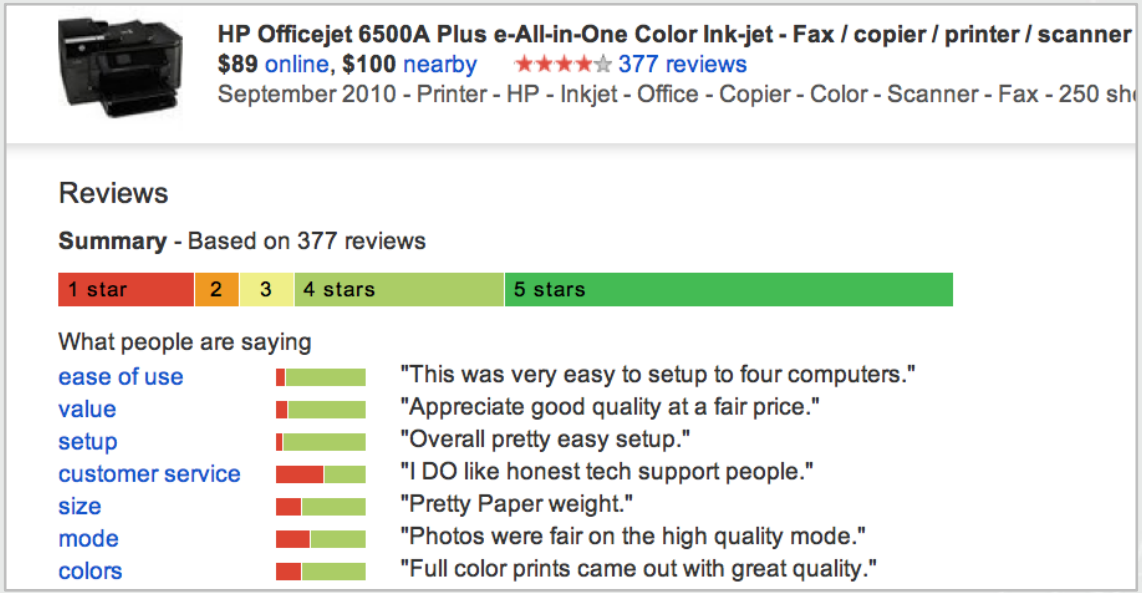
Note: The blue text marks a location name and the underlined text marks a commodity name. *mmif* is the abbreviation for Marra Mamba Iron Formation.



Associate Prof. Wei Liu & Tyler Bikaun, UWA

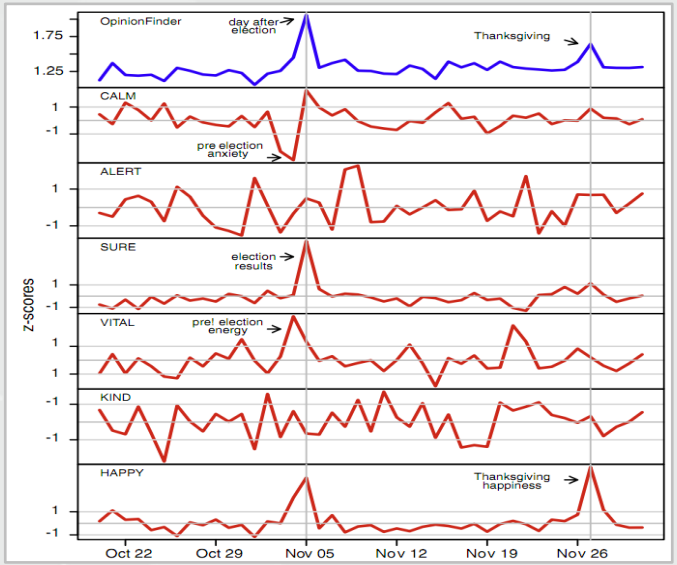
S^3 - Sentiment

Extracting Sentiment from Text

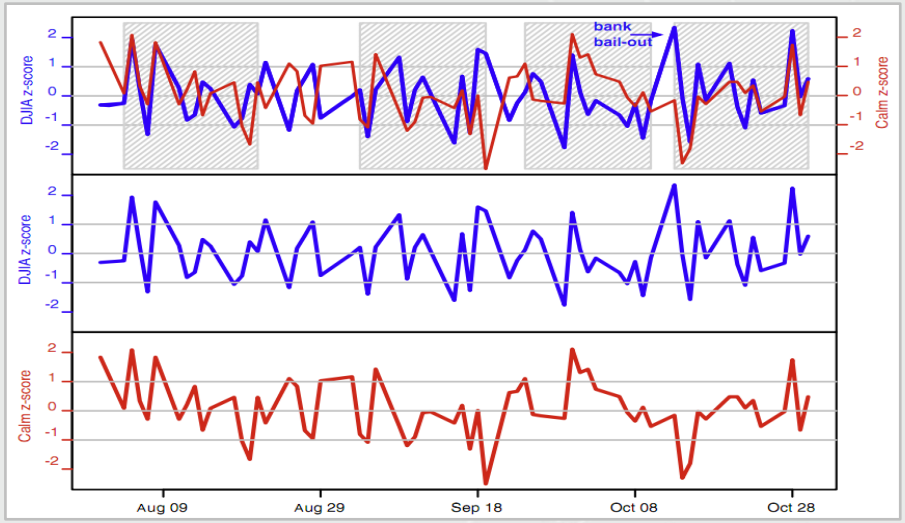


S^3 - Sentiment

Extracting Sentiment from Text



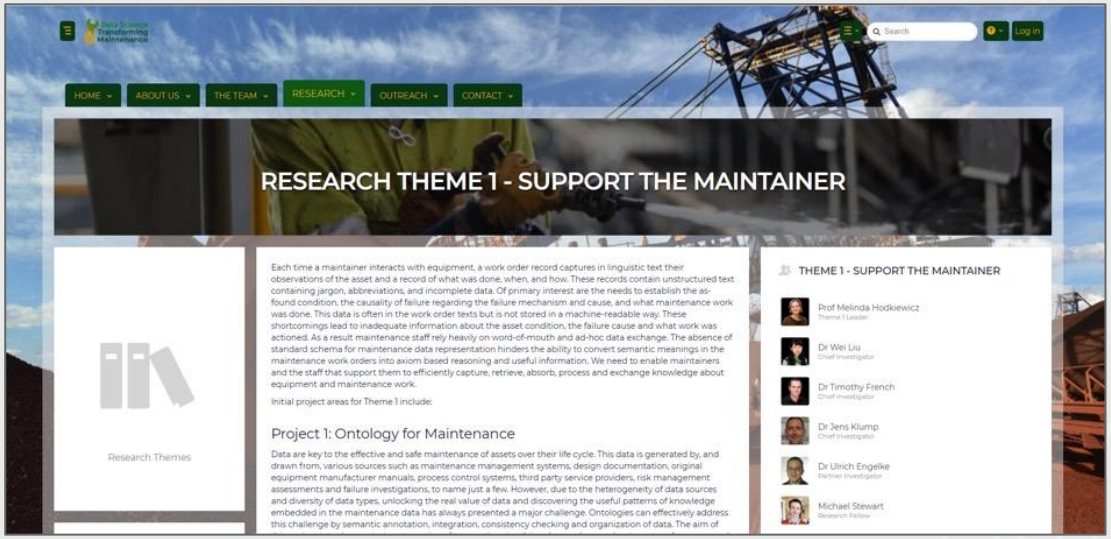
CALM Dow Jones



Associate Prof. Wei Liu & Tyler Bikaun, UWA

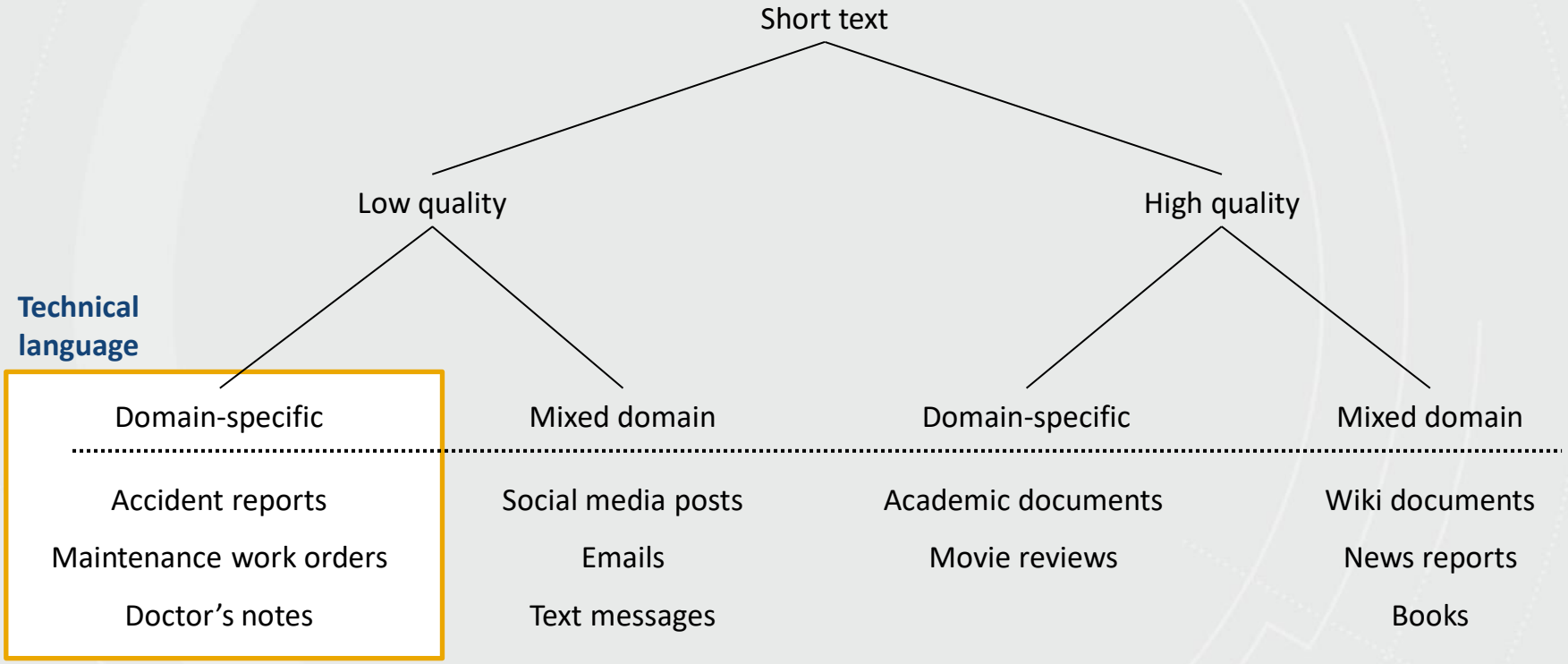
Current Projects

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



Associate Prof. Wei Liu & Tyler Bikaun, UWA

Short Text Taxonomy



Associate Prof. Wei Liu & Tyler Bikaun, UWA

Knowledge Graph Construction from Text – Text2KG Pipeline

Knowledge Graph Construction From Text

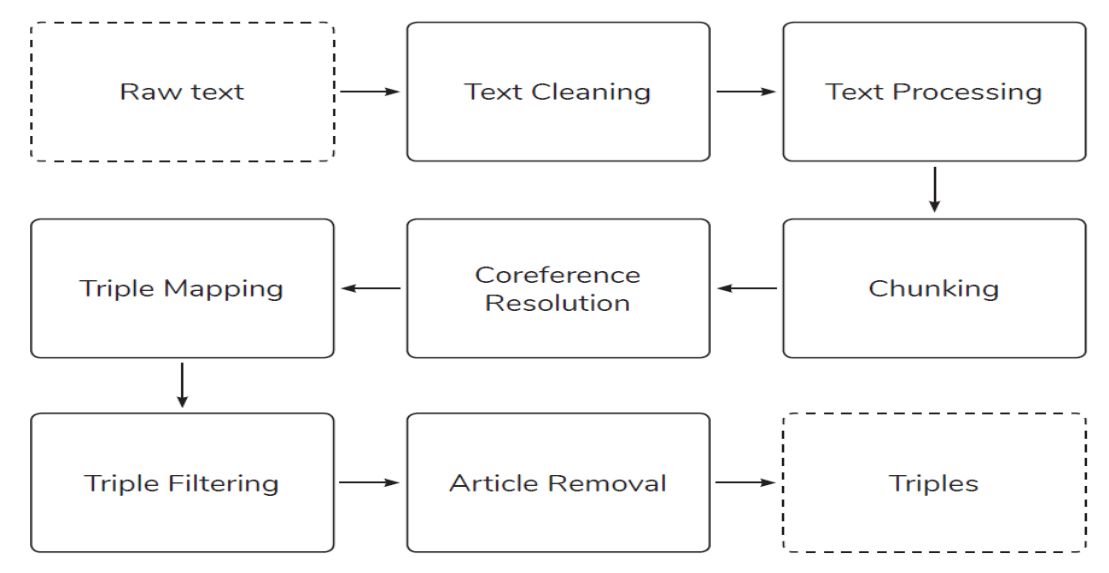
- A knowledge graph (KG) is a graph-structured knowledge base that stores knowledge in the form of the relation between entities.

- Subtasks of Knowledge Graph Construction
  - Entity Extraction
  - Relation Extraction
  - Axiom Extraction

- Note that we do not attempt to link to an existing KG such as Freebase, DBPedia, YAGO, and NELL.

Knowledge Graph Construction from Text

ICDM Knowledge Graph Contest 2019



M. Stewart, M. Enkhsaikhan and W. Liu, "ICDM 2019 Knowledge Graph Contest: Team UWA," *2019 IEEE International Conference on Data Mining (ICDM)*, Beijing, China, 2019, pp. 1546-1551.

Knowledge Graph Construction from Text

Text Cleaning and Text Processing

Text Cleaning: special characters such as hyphen and quotation marks and also break sentences joined together with no space between them.

Text processing: The text is processed through tokenisation, POS tagging, entity recognition and dependency parsing steps using SpaCy.

Token Id	Token	Entity Type	IOB	Coarse Grained POS	POS	Start	End	Dependency
0	Ford	ORG	B	PROPN	NNP	0	3	compound
1	Motor	ORG	I	PROPN	NNP	5	9	compound
2	Company	ORG	I	PROPN	NNP	11	17	nsubj
3	is		O	VERB	VBZ	19	20	ROOT
4	an		O	DET	DT	22	23	det
5	American	NORP	B	ADJ	JJ	25	32	amod
6	multinational		O	ADJ	JJ	34	46	amod
7	automaker		O	NOUN	NN	48	56	attr
8	that		O	DET	WDT	58	61	nsubj
9	has		O	VERB	VBZ	63	65	relcl
10	its		O	DET	PRP	67	69	poss
11	main		O	ADJ	JJ	71	74	amod
12	headquarters		O	NOUN	NN	76	87	dobj
13	in		O	ADP	IN	89	90	prep
14	Dearborn	GPE	B	PROPN	NNP	92	99	pobj
15	,		O	PUNCT	,	100	100	punct
16	Michigan	GPE	B	PROPN	NNP	102	109	appos
17	,		O	PUNCT	,	110	110	punct
18	a		O	DET	DT	112	112	det
19	suburb		O	NOUN	NN	114	119	dobj
20	of		O	ADP	IN	121	122	prep
21	Detroit	GPE	B	PROPN	NNP	124	130	pobj
22	.		O	PUNCT	.	131	131	punct
23	The		O	DET	DT	133	135	det
24	company		O	NOUN	NN	137	143	nsubjpass
25	was		O	VERB	VBD	145	147	auxpass
26	founded		O	VERB	VBN	149	155	ROOT
27	by		O	ADP	IN	157	158	agent
28	Henry	PERSON	B	PROPN	NNP	160	164	compound
29	Ford	PERSON	I	PROPN	NNP	166	169	pobj
30	and		O	CCONJ	CC	171	173	cc
31	incorporated		O	VERB	VBD	175	186	conj
32	on		O	ADP	IN	188	189	prep
33	June	DATE	B	PROPN	NNP	191	194	pobj
34	16	DATE	I	NUM	CD	196	197	nummod
35	,	DATE	I	PUNCT	,	198	198	punct
36	1903	DATE	I	NUM	CD	200	203	nummod
37	.		O	PUNCT	.	204	204	punct

TABLE I
TEXT PROCESSING: TOKENISATION, POS TAGGING, ENTITY RECOGNITION, AND DEPENDENCY PARSING.

Associate Prof. Wei Liu & Tyler Bikaun, UWA

Knowledge Graph Construction from Text

Phrase Identification – Linguistic Rules

Noun Phrase Chunking:

- NP
- (NP)
- NP + of + NP
- NP + NP

Verb Phrase Chunking:

- VERB + PART
- VERB + ADP
- ADP + VERB
- PART + VERB

Sent #	Phrase #	Phrase	Type
0	0	Ford Motor Company	ENTITY
0	1	is	VERB
0	2	an American multinational automaker	ENTITY
0	3	that	DET
0	4	has	VERB
0	5	its main headquarters	ENTITY
0	6	in	ADP
0	7	Dearborn	ENTITY
0	8	,	PUNCT
0	9	Michigan	ENTITY
0	10	,	PUNCT
0	11	a suburb of Detroit	ENTITY
0	12	.	PUNCT
1	13	The company	ENTITY
1	14	was founded by	VERB
1	15	Henry Ford	ENTITY
1	16	and	CCONJ
1	17	incorporated on	VERB
1	18	June 16, 1903	ENTITY
1	19	.	PUNCT

Knowledge Graph Construction from Text

Triple Creation and Filtering

COREF Resolution *: Coreference items are resolved by replacing the original phrase with the referred phrase for each item.

Triple Creation: Find triples in the same sentence by finding entities before and after verb phrases; Document-level triples are obtained by finding shortest path between any pair of entities.

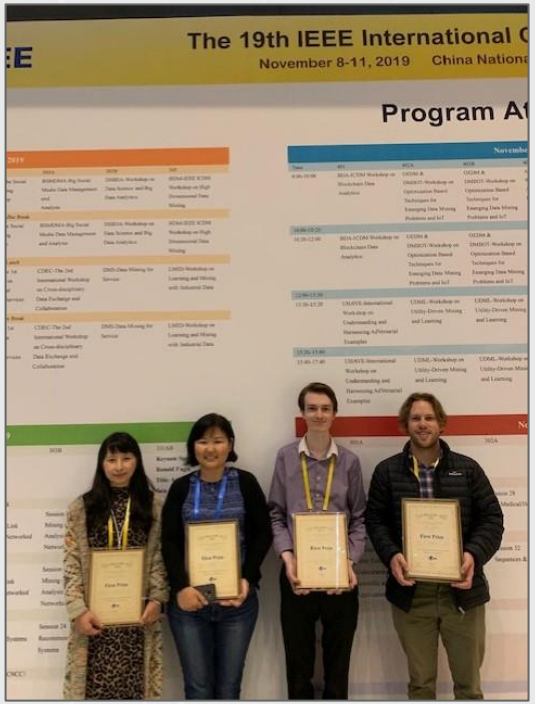
Triple Filtering: remove any triple with a stop word as a head entity. The stop words include NLTK stop words, names of days (e.g. Monday) and names of months (e.g. January).

Article Removal: articles (a, an, the), possessive pronouns (e.g., its, their) and demonstrative pronouns (e.g., that, these) are removed from the heads and tails.

Triple			Additional information						
Head ($_H$)	Relation ($_R$)	Tail ($_T$)	SemEval Relation	Type$_H$	Type$_T$	Deg$_H$	Deg$_T$	Betw$_H$	Betw$_T$
Ford Motor Company	in	Dearborn	Content-Container	ORG	LOC	6	3	11.0	0.75
Ford Motor Company	in	Michigan	Content-Container	ORG	LOC	6	3	11.0	0.75
Ford Motor Company	in	suburb of Detroit	Member-Collection	ORG	O	6	3	11.0	0.75
Ford Motor Company	in	June 16, 1903	Component-Whole	ORG	O	6	2	11.0	0.0
Ford Motor Company	is	American multinational automaker	Instrument-Agency	ORG	O	6	5	11.0	1.75
Ford Motor Company	was founded by	Henry Ford	Product-Producer	ORG	PER	6	2	11.0	0.0
American multinational automaker	in	Dearborn	Member-Collection	O	LOC	5	3	1.75	0.75
American multinational automaker	in	Michigan	Member-Collection	O	LOC	5	3	1.75	0.75
American multinational automaker	in	suburb of Detroit	Member-Collection	O	O	5	3	1.75	0.75
American multinational automaker	has	main headquarters	Cause-Effect	O	O	5	4	1.75	1.0
Henry Ford	incorporated on	June 16, 1903	Component-Whole	PER	O	2	2	0.0	0.0
main headquarters	in	Dearborn	Content-Container	O	LOC	4	3	1.0	0.75
main headquarters	in	Michigan	Content-Container	O	LOC	4	3	1.0	0.75
main headquarters	in	suburb of Detroit	Member-Collection	O	O	4	3	1.0	0.75

NeuralCoref: https://github.com/huggingface/neuralcoref

Text2KG: ICDM Knowledge Graph Contest – First Prize



Associate Prof. Wei Liu & Tyler Bikaun, UWA

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)

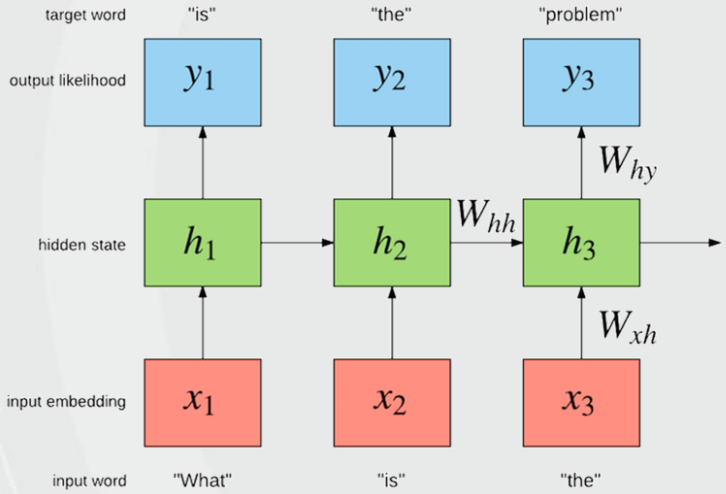
- Tokenization
- Stop-words
- Feature distributions
- N-grams
- Stemming
- Concordancing
- Dispersion plotting
- Bi-gram significance
- Word contexts
- Word similarities
- Parts-of-Speech (POS)
- Named entity recognition

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?

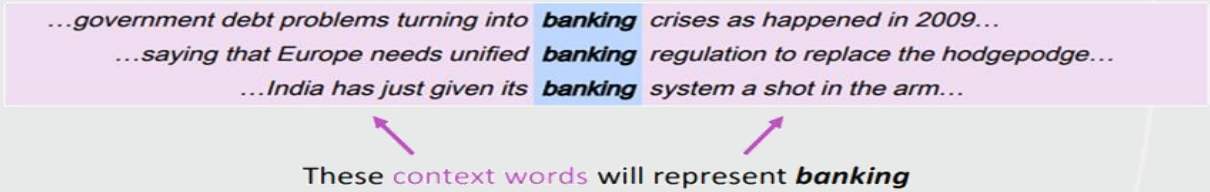
What	=	[10000000]
is	=	[01000000]
the	=	[00100000]
name	=	[00010000]
of	=	[00001000]
the	=	[00100000]
prime	=	[00000100]
minister	=	[00000010]
of	=	[00001000]
Australia	=	[00000001]

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]

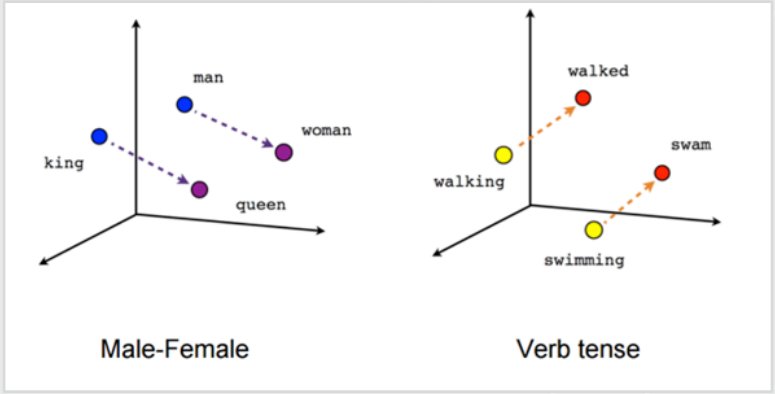
One-hot **Embeddings**

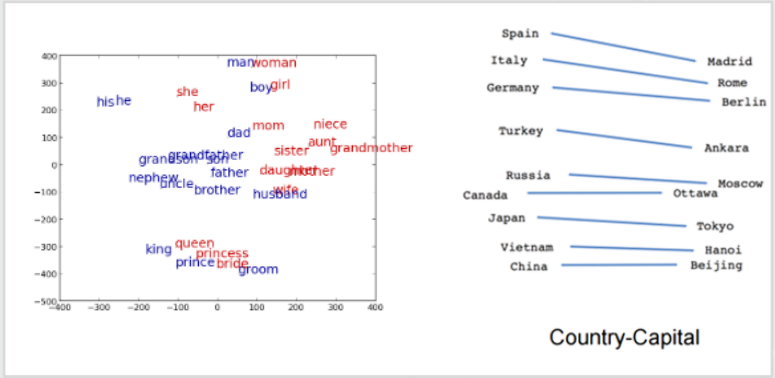
Word Embeddings: Power of Embeddings

Imagine doing simple math with words

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

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





Associate Prof. Wei Liu & Tyler Bikaun, UWA

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

Word Embeddings: Analogy Task Examples

“A is to B as C is to D” tasks

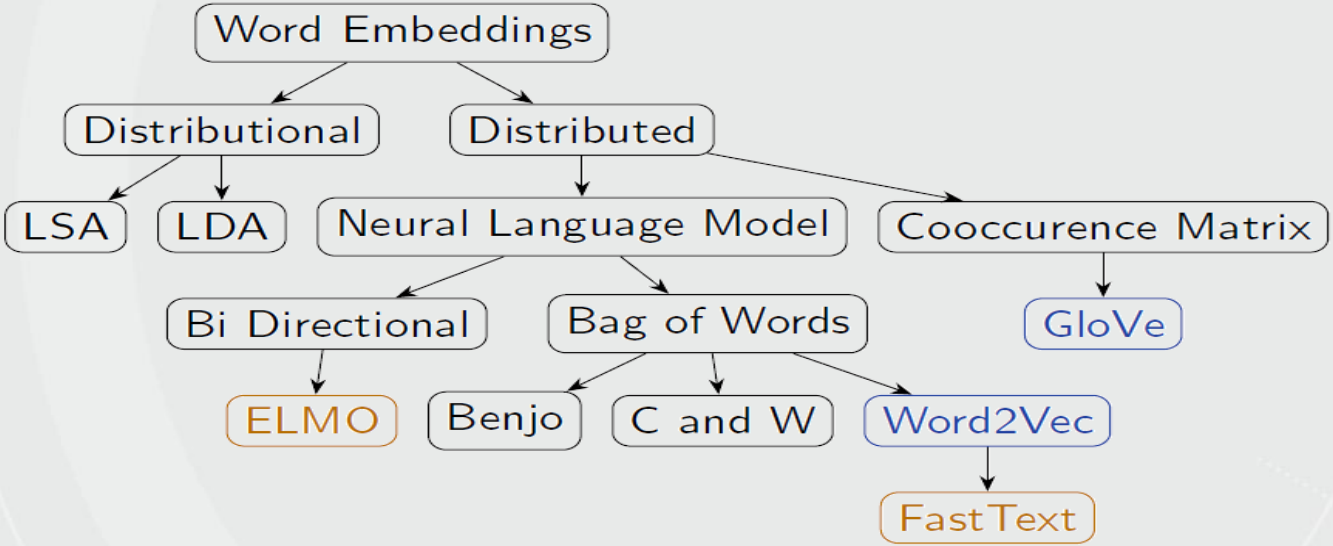
groom bride husband wife
groom bride king queen
groom bride man woman
groom bride nephew niece
groom bride policeman policewoman
groom bride prince princess

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

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)

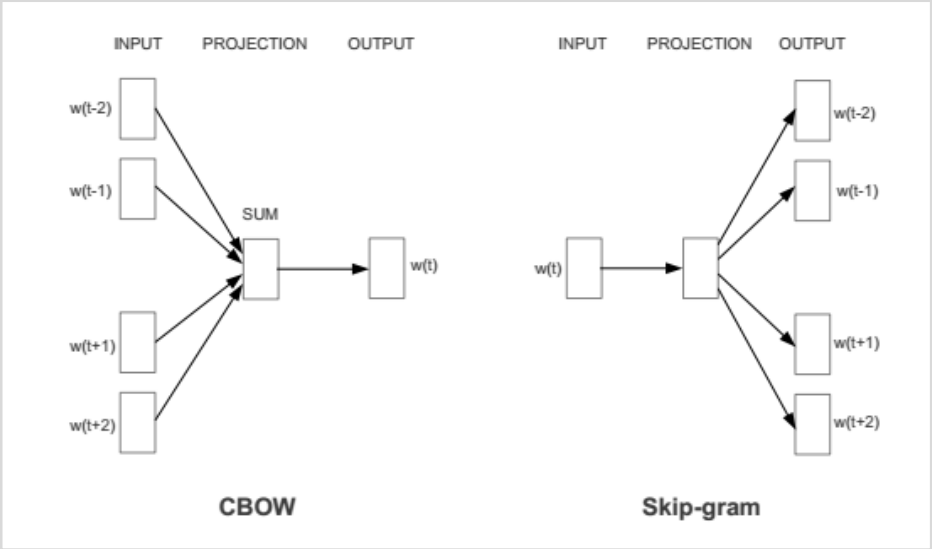
- For learning word representations using a shallow neural network
- No shared parameters across the network
- Fast training
  - Training over Wikipedia snapshot (about 1 billion words) with 50 dimensions takes only 4 hours
- Strong representations – previous state-of-the-art performance.
- Learned word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

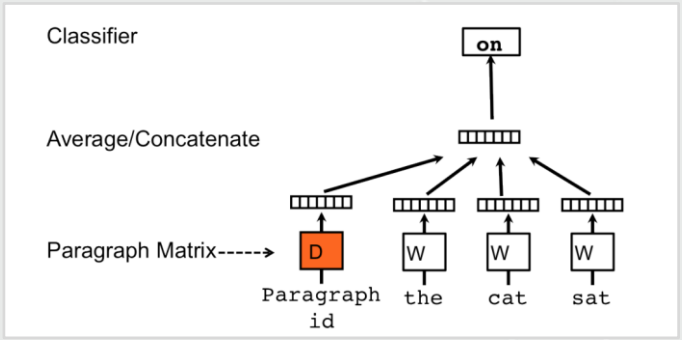


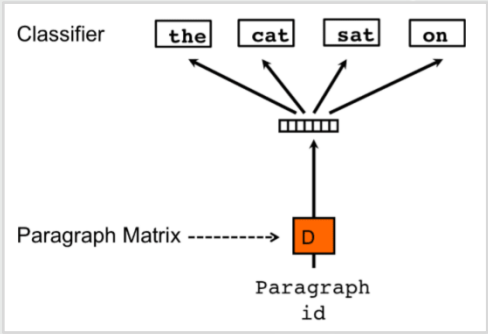
Other Language Representations

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/

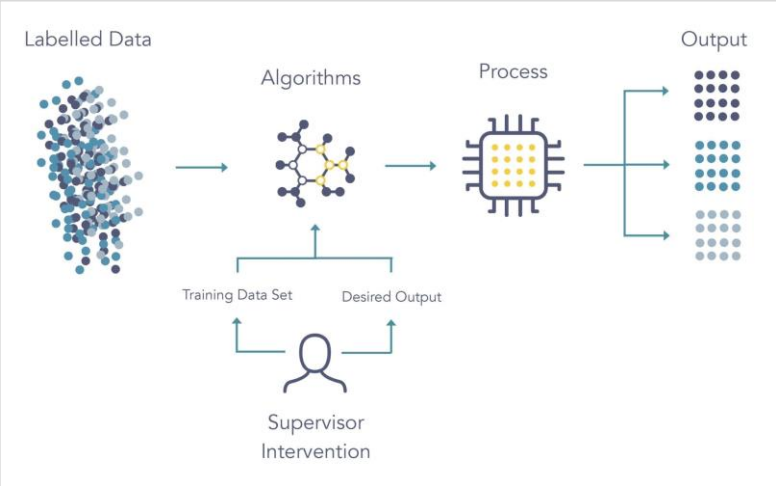
12.3 Supervised Learning

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

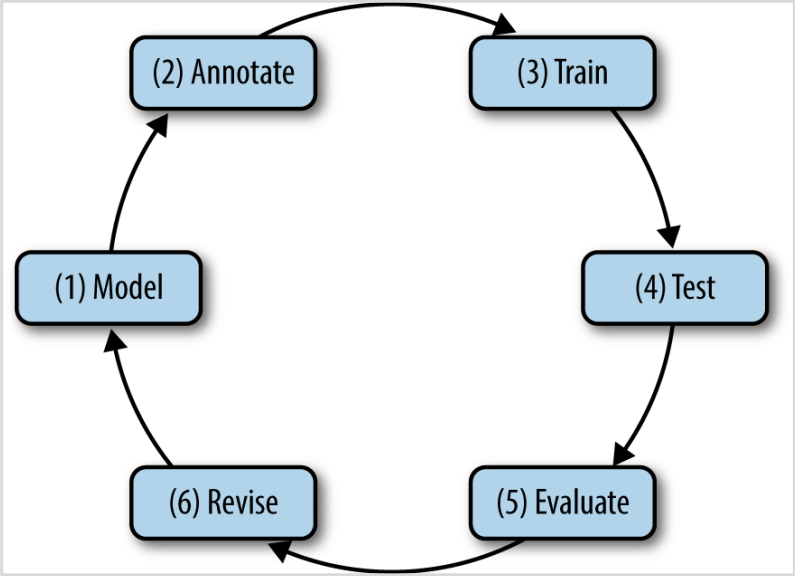
Agenda (Hands-on Session) – Exploring Supervised NLP Tasks using HuggingFace 🤗

- 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

Associate Prof. Wei Liu & Tyler Bikaun, UWA

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

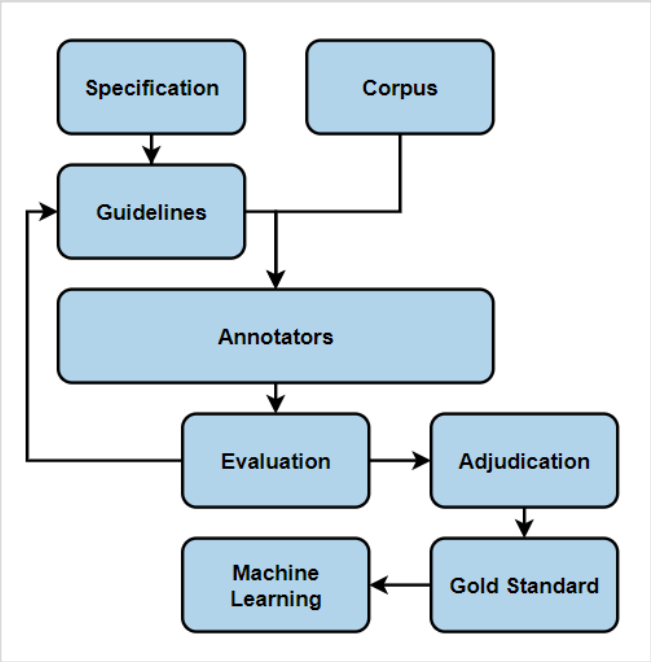
body part
body part/leg
body part/leg/ankle

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*"

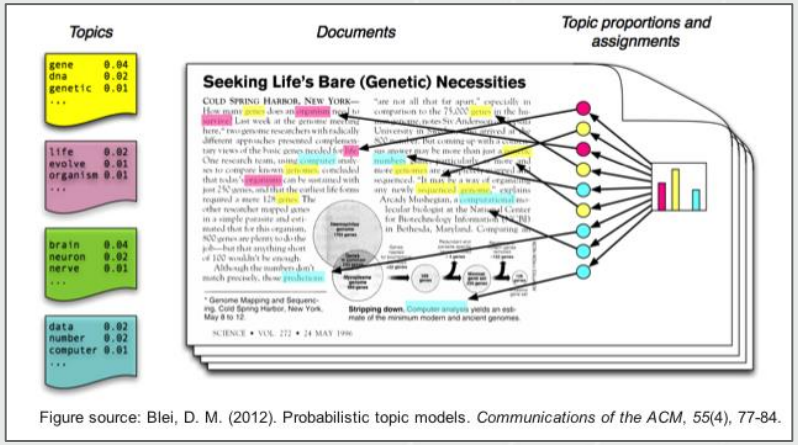


Figure source: Blei, D. M. (2012). Probabilistic topic models. *Communications of the ACM*, *55*(4), 77-84.

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

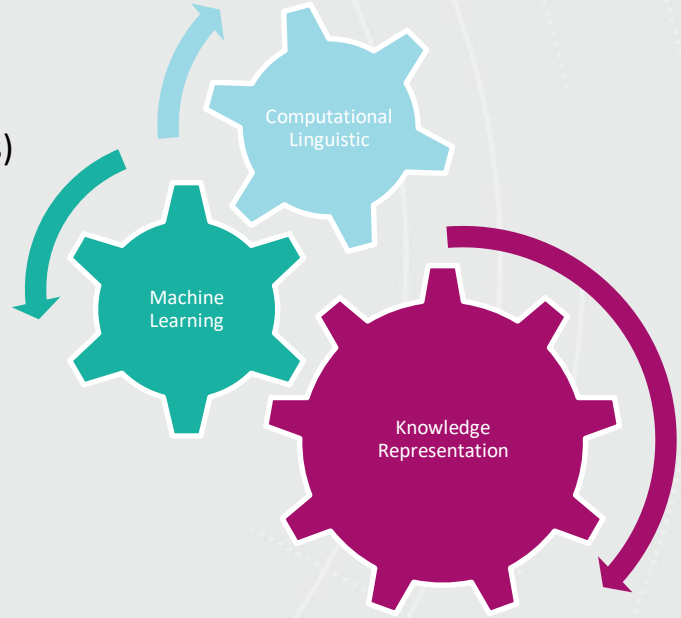
A CORE partnership

COREHUB.COM.AU/SKILLS

Additional Content

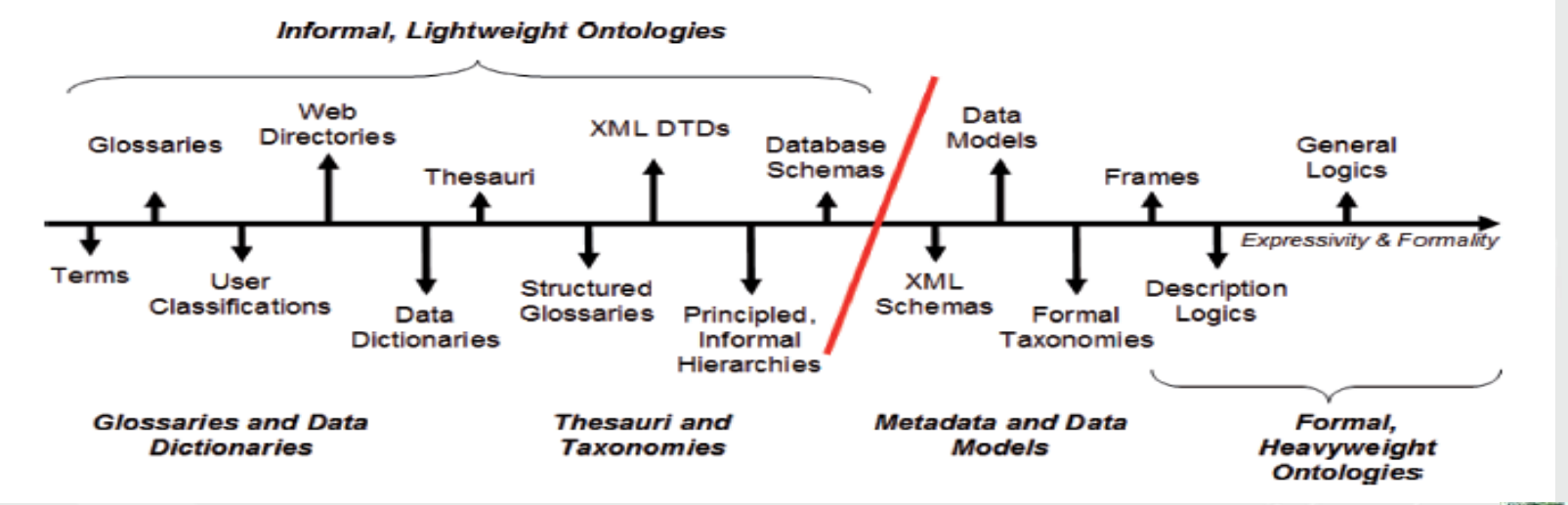
NLP Pipelines – Ontology Learning from Text

- **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...

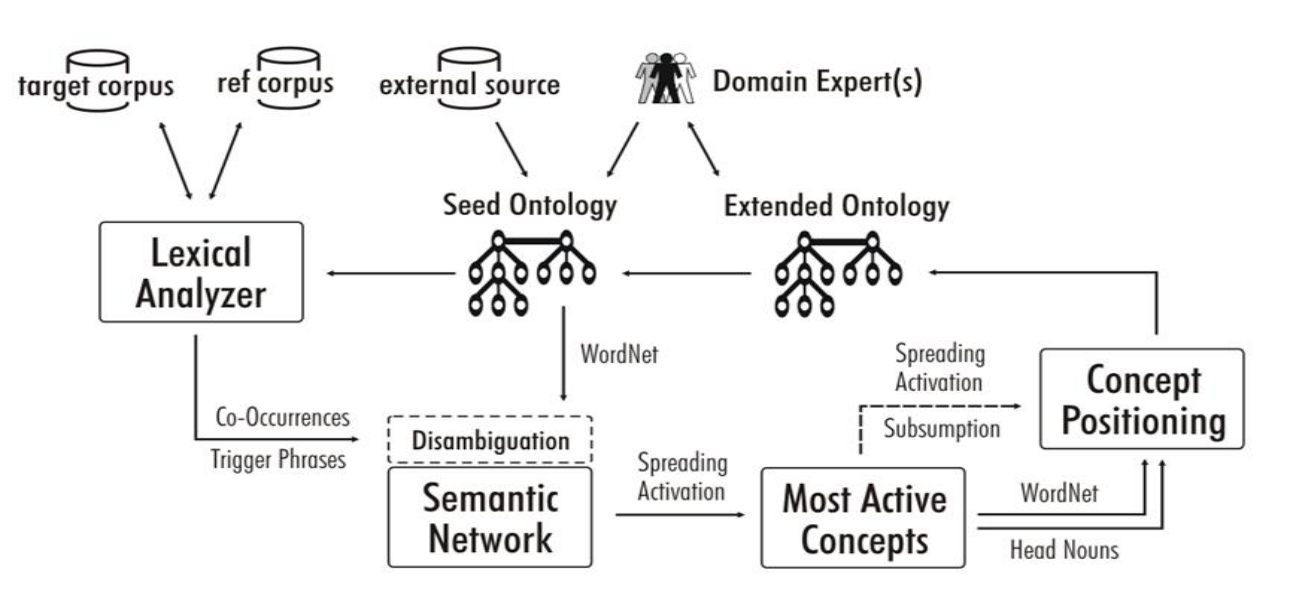
...[University of WA]...[universities] of [Perth]....

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

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

Ontology Learning from Text – System Overview

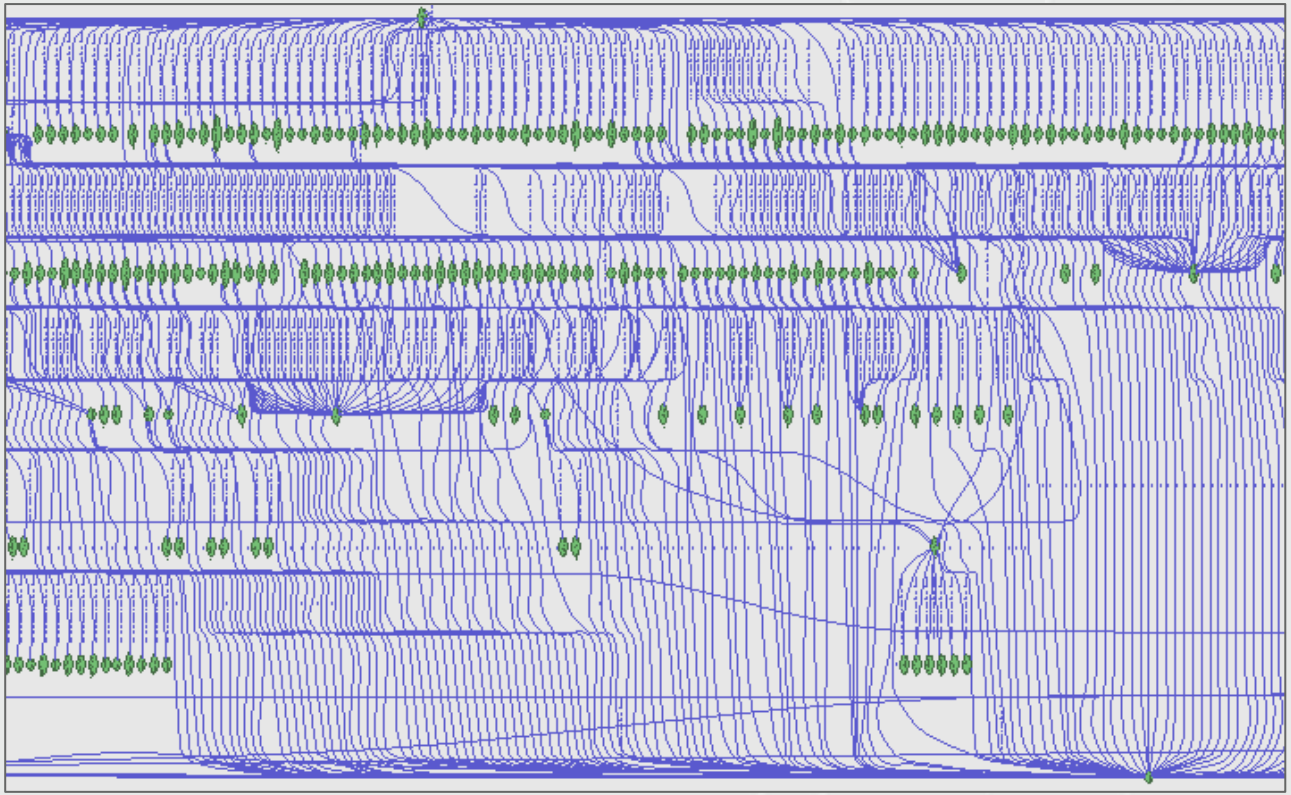
A semi-automatic ontology learning system



Wei Liu, Arno Scharl & Albert Weischselbraun, 2005

Ontology Learning from Text – Term Co-ocurrence

Semantic Network Visualisation



Ontology Learning from Text – Term Significance

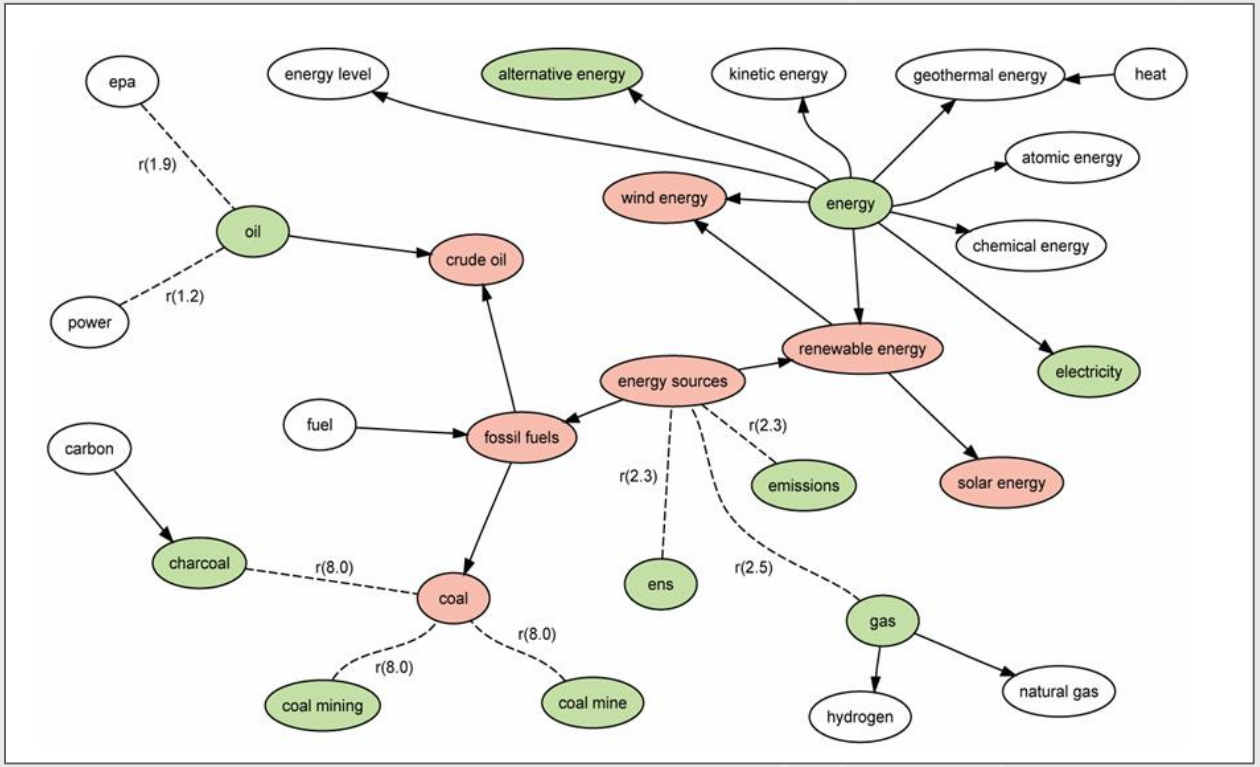
Results of climate change data analysis

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

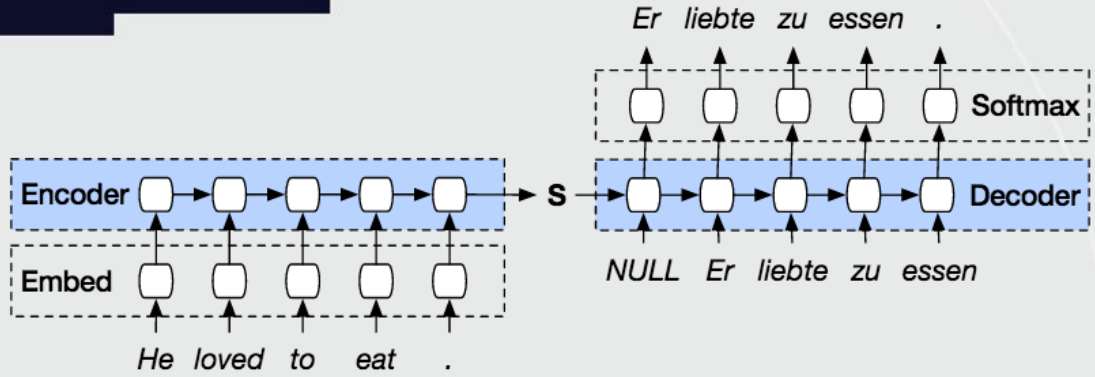
Ontology Learning from Text – Building Lightweight Ontologies

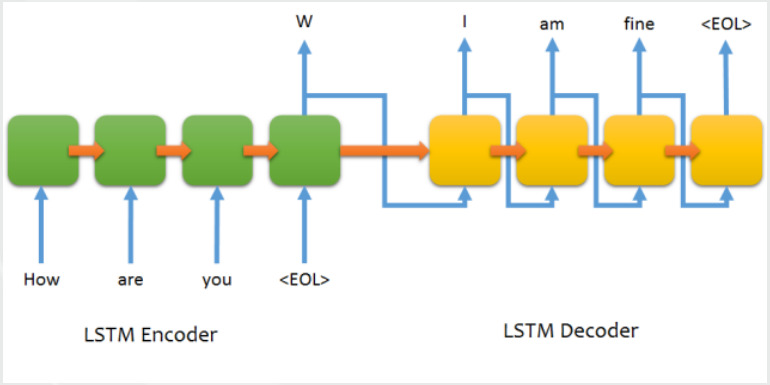
Spreading Activation

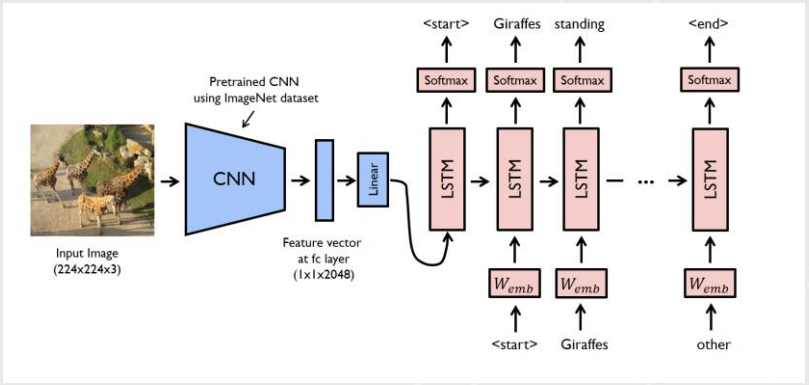


12.4 Neural Language Processing

Sequence to Sequence Learning (Encode then Decode)

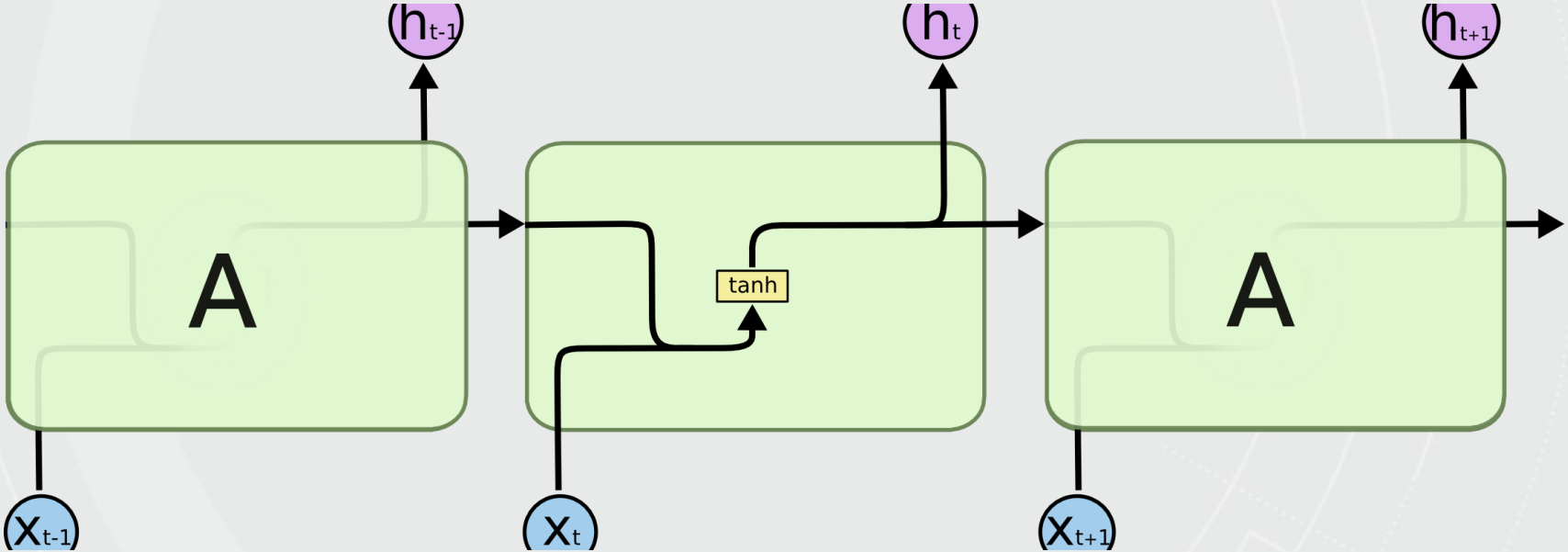






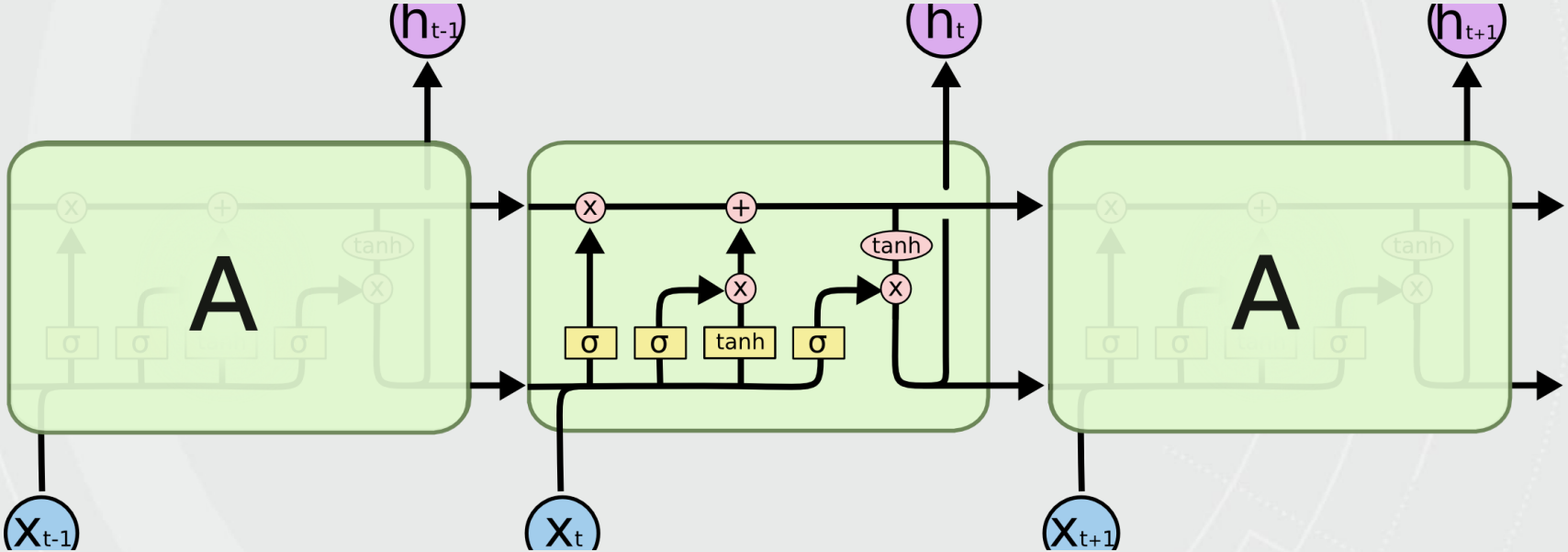
LSTM

Contains four interacting layers

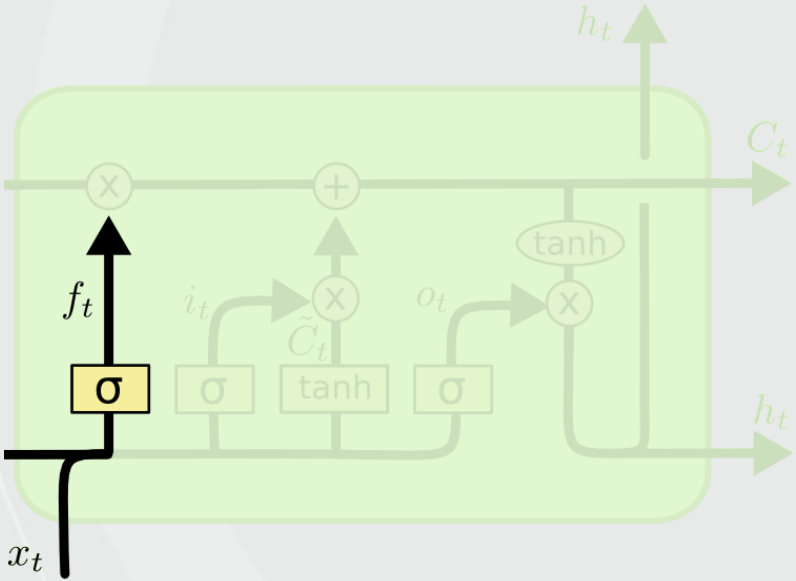


LSTM

Inside a cell



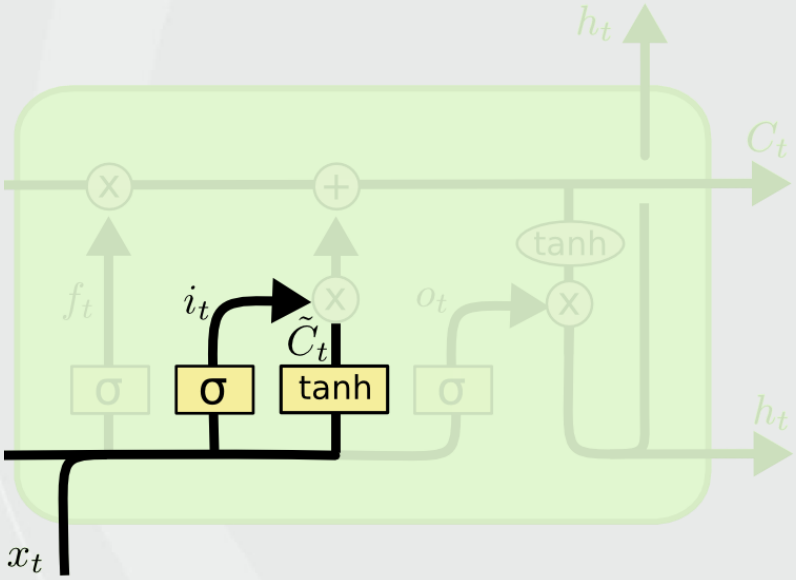
LSTM – Forget gate



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

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

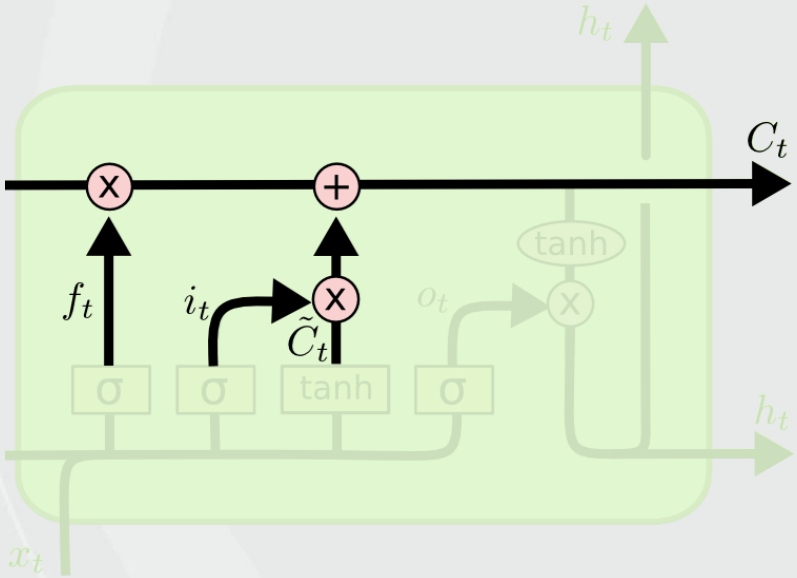
LSTM – Input Gate Layer



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

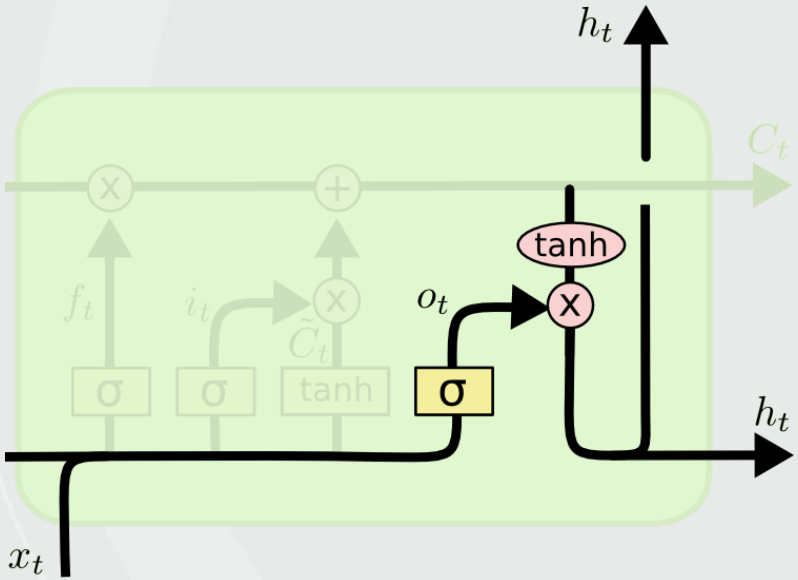
$$\tilde{C}_t = \tanh(W_C \cdot [h_{t-1}, x_t] \; + b_C)$$

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)