Natural Language and Text Analytics

Andrea Esuli

The use of a complex language is a **distinctive trait of humans**.

- Thousands of symbols.
- Complex syntax.
- Semantic is *mostly* compositional.
- Potentially **ambiguous**.
- Relies on shared world knowledge.
- Learned from scratch **along life**.
- It **evolves** along our lives.

James R Hurford, "Human uniqueness, learned symbols and recursive though" Thomas C. Scott-Phillips, Richard A. Blythe, "Why is combinatorial communication rare in the natural world, and why is language an exception to this trend?"

• Thousands of symbols.



ames or) one of e played at risks,	dictatorial /,diktetorial/ add dictatorial /,diktetorial/ add like a dictator. 2 overbearing orially adv. [Latin: related
cut into	diction /'dikf(a)n/
AND BID	ciation in speaking or singing dictio from dico dict- say
ricky	alctionary /dikjanari/ n
es) di- efined	explaining the words of a lan giving corresponding words in language. 2 reference book
ed to	the terms of a particular

• Complex syntax.





• Semantic is *mostly* compositional.

compositional:

to buy a car leggere un libro

collocation:

make the bed saltare la lezione

idiomatic expression:

break a leg dalla padella nella brace

- Potentially ambiguous.
 - Part of speech ambiguity

beat: Verb, Noun? *Listen to this cool beat. I'll beat you at checkers!*

• Word sense ambiguity

interest: attention or money? *There is interest on this topic. The bank raised the interest rate.*

 \circ Syntactic ambiguity \rightarrow



• Relies on shared world knowledge.

"I'm faster than Lewis Hamilton"

"Alexis and Kate are sisters"

"Alexis and Kate are mothers"

"My mother is younger than me."

"I was on a bike with a helmet"

"I was on a bike with an electric engine"

Faster doing what?

Relation.

Same structure, no relation.

Not possible.

The helmet is on my head.

The engine is part of the bike.

• Learned from scratch along life.



30 million words gap

A <u>study</u> started in 1982 (<u>currently</u> <u>debated</u>) by Hart and Risley measured the number of words listened from their parent by children during the first three years of their life, when the main source of experience about the world is the communication with the parents.

They observed a gap of 30 million words (a factor of four) between children living in high income families and those living in low income families.



30 million words gap

The study than found that the measure of number of word listened from parent in the early years is a good predictor of the child future language skills, even at distance of years.

"... We were awestruck at how well our measure of accomplishments at age 3 predicted measures of language skill at age 9-10 ... Vocabulary use at age 3 was equally predictive of measures of language skill at age 9-10 ..."

<u>Successive studies</u> measured different gaps, and discovered significant biases in the original study.

Yet, many studies confirmed the positive effect of experiencing more use of language in the early, **pre-scholar** phases of life.

• It evolves along our lives.

"<u>airplane</u>"

"<u>quark</u>"

"<u>automobile</u>"

"<u>genocide</u>"

"Tesla" (<u>books</u>, <u>web</u>)

"a <u>ship</u>"

"I googled it"

"it's in the cloud"

"<u>crowdfunded</u>"

"to navigate"

"super spreader"

"wear a mask"

The use of a complex language is a **distinctive trait of humans**.

- Thousands of symbols.
- Complex syntax.
- Semantic is *mostly* compositional.
- Potentially ambiguous.
- Relies on shared world knowledge.
- Learned from scratch along life.
- It evolves along our lives.

James R Hurford, "Human uniqueness, learned symbols and recursive though" Thomas C. Scott-Phillips, Richard A. Blythe, "Why is combinatorial communication rare in the natural world, and why is language an exception to this trend?"

Natural Language Understanding (by machines)

Natural Language Understanding aims at building machines that are able to receive and give information using natural language, like humans do.

From a computational point of view, natural language understanding is considered to be an <u>Al-complete</u> problem.

Practical text analytics tasks are simplifications of NLU that make the problem easier to solve.



"Text mining is the **analysis of data contained in natural language text**. Text mining works by transposing words and phrases in **unstructured data** into numerical values which can then be linked with **structured data** in a database and analyzed with **traditional data mining techniques**."



Text mining supports data analysis problems by **recognizing** and **extracting structured knowledge** that is otherwise locked in the **unstructured** form of **natural language**, and thus **out of reach** from traditional **data mining** methods.

A complete, integrated text mining and data mining pipeline can be faster than humans at taking <u>critical decisions</u>.

A	Symbol	Open	High	Low	Close	Net Chg	%Chg	Vol
A10 Networks	ATEN	7.72	7.88	7.59	7.85	0.14	1.82	462,746
AAC Holdings	AAC	24.62	24.82	21.14	24.72	0.18	0.73	149,108
AAR Corp.	AIR	24.61	24.72	24.42	24.65	0.09	0.37	114,148
Aaron's Inc.	AAN	24.42	24.75	24.27	24.50	0.23	0.95	808,727
ABB ADR	ABB	18.74	18.78	18.48	18.60	-0.22	-1.17	2,599,799
Abbott Laboratories	ABT	45.22	45.52	44.94	45.46	0.54	1.20	5,522,257
AbbVie	ABBV	58.23	59.09	57.16	59.02	0.87	1.50	11,025,268
Abercrombie&Fitch	ANF	25.90	26.32	25.70	25.75	0.18	0.70	4,881,104
ABM Industries	ABM	29.75	30.13	29.59	30.10	0.43	1.45	168,706
Acadia Realty Trust	AKR	33.41	33.94	33.01	33.55	0.01	0.03	302,163
Accenture CI A	ACN	107.29	108.25	107.22	108.13	0.91	0.85	2,405,340
ACCO Brands	ACCO	7.76	7.88	7.70	7.88	0.16	2.07	358,227
Accuride	ACW	2.57	2.57	2.45	2.50	-0.08	-3.10	84,018
ACE	ACE	114.63	116.22	114.54	116.13	1.28	1.11	1,811,121
Acorn International ADR	ATV	2.73	20.85	2.62	19.20	16.73	677.49	5,070,208
Actuant CI A	ATU	24.79	24.96	23.93	24.36	-0.40	-1.62	623,034
Acuity Brands	AYI	232.17	233.74	231.60	233.22	2.34	1.01	214,090
Adecoagro	AGRO	11.28	11.28	10.84	10.98	-0.27	-2.40	307,478

Nokia Launches \$16.6 Billion Offer for Alcatel

Nokia CEO Rajeev Suri says merger will mean 'we are just ahead of the curve'

By JENS HANSEGARD

Nov. 18, 2015 6:19 a.m. ET

HELSINK1—Finnish wireless-equipment specialist Nokia Corp. on Wednesday commenced its share-exchange offer for Alcatel-Lucent SA shareholders in Paris and New York, betting its proposed €15.6 billion (\$16.6 billion) acquisition will allow a combined company to better compete in a global race for scale in the business of making telecommunications and Internet gear.

Nokia's attempt to create a one-stop shop for telecom companies and Internet service providers comes as the company faces heightened competition from new...

TM methods may be aimed at **directly extracting** the answer to an information need from text,

or

give an **explicit structure** to the information contained in text in order to enable the application of data mining methods



accuracy algorithm applications automatically based class classification classifier codes data dataset different documents estimating evaluation experiments expressed feature functions generated given human improve information inspection labelled learning measure methods opinion optimized positive present problem product proposed quantification ranking reports results review selection standard strategy supervised system task text training used

85 percent of business-relevant information originates in unstructured form, primarily text, according to Anant Jhingran of IBM Research.

Information is mostly communicated by reading or writing e-mails, reports, or articles and the like, in conversations, or by listening/watching media

Attempts to turn text into (semi-)structured forms (HTML) or <u>microformats</u> (atom, card, calendar, geo) only scratch the surface, but:

- Requires universal agreed ontologies
- Additional effort to make structure explicit

Entity linking attempts to provide a bridge.

The vision...



...the reality

IF HAL-9000



WAS ALEXA

Text Mining Tasks

The are many different practical information processing tasks in which text is the main media:

- Tagging
- Parsing
- Search
- Question Answering
- Summarization
- Translation
- Classification
- Clustering

- Regression
- Quantification
- Information Extraction
- Named Entity Recognition
- Coreference Resolution
- Entity linking
- Paraphrase
- Dialog

Text Mining Tasks

The performance of the available methods of the tasks varies:

Very good:

- Tagging
- Parsing
- Named Entity Recognition
- Search

Progressing:

- Translation
- Classification
- Clustering
- Coreference Resolution

- Regression
- Information Extraction
- Entity linking

Not yet ready:

- Question Answering
- Summarization
- Quantification
- Paraphrase
- Dialog

Text mining methods find application in many everyday situations.



Q Cerca su Twitter

Mostra altro

Tendenze - Italia	ලා
1 · Di tendenza #autocertificazione 1.615 Tweet	~
2 · Di tendenza #Ratzinger 6.162 Tweet	~
3 · Di tendenza #Pompeo 1.832 Tweet	~
4 · Di tendenza #domenicain	~
5 · Di tendenza #Hunziker 8.072 Tweet	~

Tendenze - Spagna	503
1 · Di tendenza Obtuve 12 de 13 1.864 Tweet	~
2 · Di tendenza #EscapeRoomLDDSS	~
3 · Di tendenza #VivaLaVida294 1.615 Tweet	~
4 · Di tendenza #liarlapardo82	~
5 · Di tendenza Alvise 18.200 Tweet	~
Mostra altro	

S

Tendenze - Giappone	ŝ
1 · Di tendenza #日向坂で会いましょう 26.300 Tweet	~
2.Di tendenza #乃木坂工事中 38.500 Tweet	~
3 · Di tendenza #関ジャム 27.900 Tweet	~
4 · Di tendenza #未来少年 コナン 22.900 Tweet	~
5 · Di tendenza #京本大我入所 14 周年 13.900 Tweet	~
Mostra altro	



Step 2: Create a Test Directory and Enable SimpleHTTPServer

3. Create a test directory where you don't mess with system files. In my case I have a partition called *[xol]* and I have created a directory called *tecmint* in there and also I have added some test files for testing.



Create Testing Directory

4. Your prerequisites are ready now. All you have to do is try python's **SimpleHTTPServer** module by issuing below command within your test directory (In my case, /x01//).



Wine 1.8 Released After 17 Months of Development – Install on RHEL/CentOS and Fedora

190

F

99

8.

44

in

Q11

COMMENTS

Install Apache 2.2.15, MySQL 5.5.34 & PHP 5.5.4 on RHEL/CentOS 6.4/5.9 & Fedora 19-12

WHITE HAT HACKER BUNDLE

DOWNLOAD FREE LINUX EBOOKS

- Complete Linux Command Line Cheat Sheet
- The GNU/Linux Advanced Administration Guide
- Securing & Optimizing Linux Servers
- Linux Patch Management: Keeping Linux Up To Date
- Introduction to Linux A Hands on Guide
- Understanding the Linux® Virtual Memory Manager
- Linux Bible Packed with Updates and Exercises
- A Newbie's Getting Started Guide to Linux
- Linux from Scratch Create Your Own Linux OS

≡ Google Translate		9
🗙 Text Documents		
JAPANESE - DETECTED JAPANESE ITAL ✓ ↔	ITALIAN ENGLISH JAPANESE	~
安倍晋三首相は、5月31日までの約1か月ま × でに日本の緊急事態を延長し、新しいコロ ナウイルスと戦うことを計画していると政 府当局者が日曜日に言った。 Abe shinzō shushō wa, 5 tsuki 31-nichi made no yaku 1- kagetsu made ni Nihon no kinkyū jitai o enchō shi, atarashī koronauirusu to tatakau koto o keikaku shite iruto seifu	Il primo ministro Shinzo Abe ha in programma di estendere l'emergenza del Giappone di circa un mese entro il 31 maggio e combattere il nuovo coronavirus, ha detto domenica un funzionario del governo.	
<u>Show more</u> ↓ ↓) 72/5000 ✓	۹)	• •

Emmanuel Macron 🔗

C'est un moment déterminant pour la communauté mondiale. En nous mobilisant aujourd'hui autour de la science et de la solidarité, nous semons les graines d'une plus grande unité demain.

...

This is a defining moment for the world community. By mobilizing ourselves today around science and solidarity, we will grow the seeds of greater unity tomorrow.

🌣 · Hide original · Rate this translation

Pour une mobilisation mondiale contre le virus.

« La chance ne sourit qu'aux esprits bien préparés ». Ces mots sont ceux de Louis Pasteur, l'un des plus grands scientifiques au monde, à l'origine de découvertes et avancées majeures qui ont sauvé des millions de vies a...

See More







Relations with other data types: tabular data

Tabular data usually contain numerical, categorical, or symbolic values that have explicit and fixed structure, constraints and relations.

Extracting structured information from <u>historical text sources</u> may enable data mining on information not originally designed for such processes.

Using natural language questions to query databases:

able				Example questions				
Name	No. of	Combined	#	Question	Answer			
	rengins uays	uays	1	Which wrestler had the most number of reigns?	Ric Flair			
Lou Thesz	3	3,749						
Ric Flair	8	3,103	2	Average time as champion for top 2 wrestlers?	AVG(3749,3103)=3426			
Harley Race	7	1,799						
Dory Funk Jr.	1	1,563	3	How many world champions are there with only	COUNT(Dory Funk Jr.,			
Dan Severn	2	1,559		one reign?	Gene Kiniski)-2			
Gene Kiniski	1	1,131	4	What is the number of reigns for Harley Race?	7			
	Name Lou Thesz Ric Flair Harley Race Dory Funk Jr. Dan Severn Gene Kiniski	NameNo. of reignsLou Thesz3Ric Flair8Harley Race7Dory Funk Jr.1Dan Severn2Gene Kiniski1	NameNo. of reignsCombined daysLou Thesz33,749Ric Flair83,103Harley Race71,799Dory Funk Jr.11,563Dan Severn21,559Gene Kiniski11,131	NameNo. of reignsCombined days#Lou Thesz33,7491Ric Flair83,1032Harley Race71,7993Dory Funk Jr.11,5633Dan Severn21,5594Gene Kiniski11,1314	NameNo. of reignsCombined days#QuestionLou Thesz33,74911Ric Flair83,10322Average time as champion for top 2 wrestlers?Harley Race71,79911Dory Funk Jr.11,56334Dan Severn21,5594How many world champions are there with only one reign?Gene Kiniski11,1314			

Relations with other data types: graphs

Social networks are an important source of always up-to-date information about social groups, events and debated topics.

Methods based on **graph analysis** enables discovering the **structure and the dynamics** of a network at various levels: local relations, communities, large scale behaviors.

Enriching these methods with the analysis of the actual content that is shared on a network enables to better characterize the elements of the network and to give more detailed meaning to the interactions, enabling to investigate complex processes, e.g., abnormal behaviors, manipulation attempts...



- Images and video represent a signal that is directly linked with a low-level perception of the world, i.e., vision.
- The entities involved in typical image processing tasks are usually bound to real world physical objects and actions (e.g., object tracking in autonomous driving).

This properties make vision-based problem relatively easier than language ones, allowing NN-based methods to achieve human-like performances.

Localize

Abnormalities

("heatmap")

Corona Score

Computation &

D Visualization



Cross-media processing:

• Image captioning is the task of describing the content of an image.

It has application in retrieval processes.

It is of help for people with vision problems.

A person riding a motorcycle on a dirt road.





A herd of elephants walking across a dry grass field.



Two dogs play in the grass.



Two hockey players are fighting over the puck.



A close up of a cat laying on a couch.



A skateboarder does a trick



A little girl in a pink hat is blowing bubbles.



A red motorcycle parked on the



A dog is jumping to catch a



A refrigerator filled with lots of food and drinks.



A yellow school bus parked



Describes without errors

Describes with minor errors

Somewhat related to the image

Unrelated to the image

Cross-media processing:

• Abstract visual representation of text enables smarter retrieval methods.



Cross-media processing:

The co-occurrence of text and images supports **transfer** learning, e.g., performing sentiment analysis on images.







Groundtruth: NEG, Prediction: NEU



Groundtruth: NEG Prediction:



Groundtruth: NEU, Prediction: NEG



Groundtruth: POS. Prediction: NEG



Groundtruth: NEU, Prediction: NEU





Groundtruth: NEU, Prediction: POS



Groundtruth: POS. Prediction:

Implementing Text Analytics methods

Text Analytics solutions are usually built combining methods from three disciplines:

- Natural Language Processing (NLP): studying how to recognize and process the elements and structures that build language as we humans define them.
- Information Retrieval (IR): studying the statistical properties of language, treating it as a signal on a channel, using Information Theory method.
- Machine Learning (ML): exploiting statistical/neural learning methods on properly processed representations of language to build descriptive/predictive models of some properties of interest.

Three interlocked disciplines

These disciplines are not one distinct from the other.

- NLP methods are often built themselves on top of IR and ML methods.
- ML adopts IR measures to define the goals of the learned models.
- ML often assumes language to be manipulated by NLP and IR methods to be able to work on it.



Information through pipelines

The concept of a **processing pipeline** is recurring in text analytics methods, as text is rarely processable *as is* by ML methods and it requires to be transformed into representation that are fit to the specific ML method in use.

Such transformations are not just simple conversions of format, they may actively filter and modify the information in order to be better exploited by successive stages of the pipeline, so as to improve the final outcome of the process.



The Neural Network revolution

Neural Networks had a strong comeback starting around 2000.

Three key factors made it possible:

- Lots of data from new models of data production and sharing.
- Powerful parallel computing that is well fit for NN algorithms
 - GPU, TPU...
- Improved and novel NN algorithms and architectures
 - There was no "Deep" in NN before.
 - Few researchers brought NN research through the <u>Al winter</u>...
 - to the AI renaissance: LSTM, CNN, ReLU, Dropout, Residual Networks, Attention, Capsule Networks, Transformers, GAN, RL...







Neural Networks and Language

The NN revolution had a great impact of some problems:

- Translation
- Summarization
- Question Answering



mainly due to the ability of NN to process huge amounts of text and the large size of their models, i.e., **to learn very large language models.**

Yet, NN for NLP have <u>limitations</u>, specially in areas that require <u>system 2 type</u> reasoning: e.g., paraphrase.