Computational Lexicography: from dictionaries to deep neural networks, and back

German Rigau <german.rigau@ehu.eus>


ORIGINAL MOTION PICTURE SOUNDTRACK
M-G-M PRESENTS A STANLEY KUBRICK PRODUCTION
2001: a space odyssey

2001: A SPACE ODYSSEY

KNOWLEDGE IS POWER

Foreword

“Cuando creíamos que teníamos todas las respuestas, de pronto, cambiaron todas las preguntas.”

- Mario Benedetti

"When we thought we had all the answers, suddenly, they changed all the questions. "

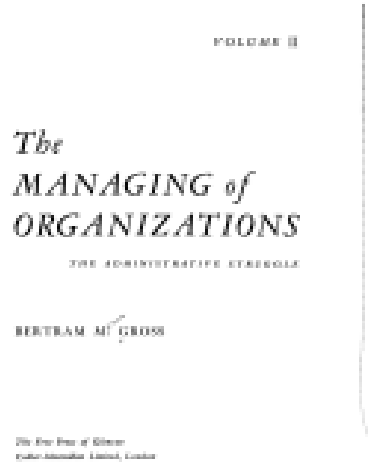
- Mario Benedetti

Foreword

- Where are the **answers** to the new (and old) questions?
  - Introspection? Experts?…
  - From many people? … “Wisdom of the Crowd”?
  - Books, News, Tweets, … Textual Sources?
  - Multimedia sources? Images, Radio, TV ...
  - Sensors? IoT? ...
  - Anything? Everything?
- Information **overload** …

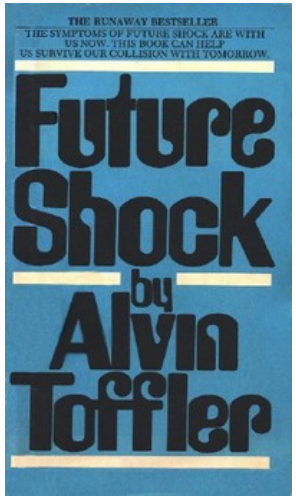
Foreword

- Information **overload**
  - infobesity, infoxication!
  - by Bertram Gross, The Managing of Organizations: The administrative struggle (1964)



Foreword

- Information **overload**
  - infobesity, infoxication!
  - by Bertram Gross, The Managing of Organizations: The administrative struggle (1964)
  - by Alvin Toffler, *Future Shock* (1970)



Foreword

- Information **overload**
  - infobesity, infoxication!
  - by Bertram Gross, The Managing of Organizations: The administrative struggle (1964)
  - by Alvin Toffler, *Future Shock* (1970)
  - Seneca complained that "*the abundance of books is distraction*" in the 1st century AD!

Foreword

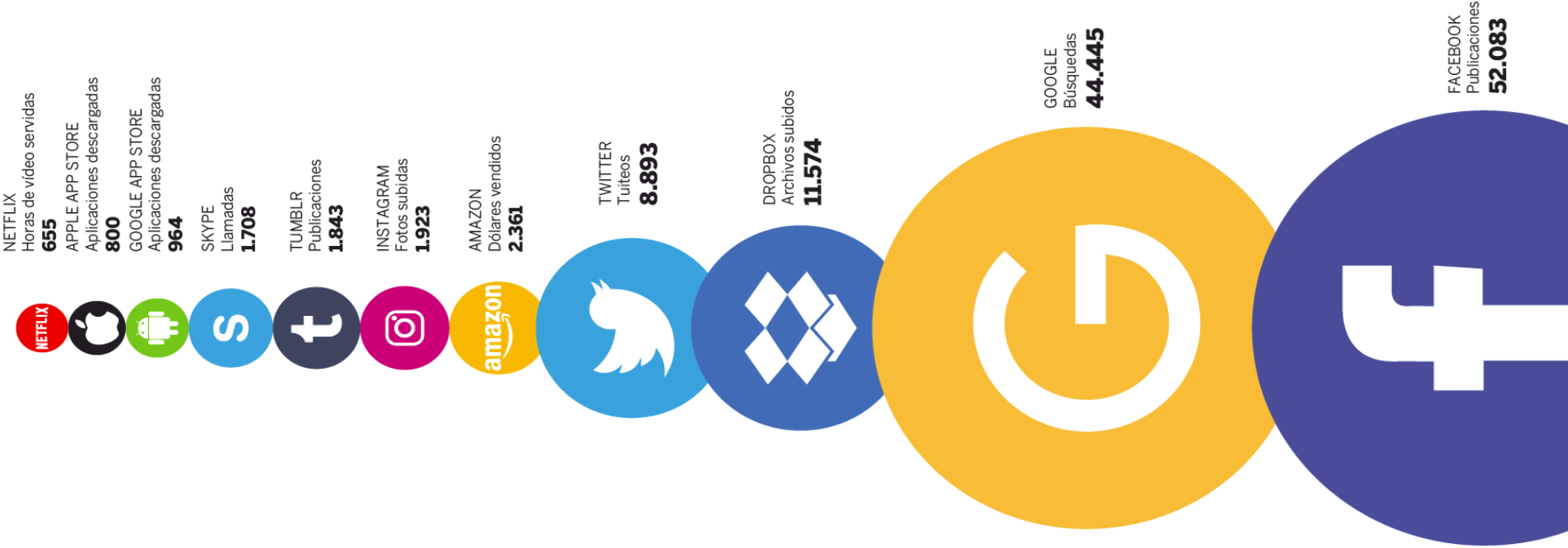
- **Information overload** occurs when the amount of input to a system exceeds its processing capacity.
- Decision makers have fairly **limited** cognitive processing capacity.
- Consequently, when information overload occurs, it is likely that a **reduction** in decision quality will occur.

- Always when **advances in technology** have increased a production of information.

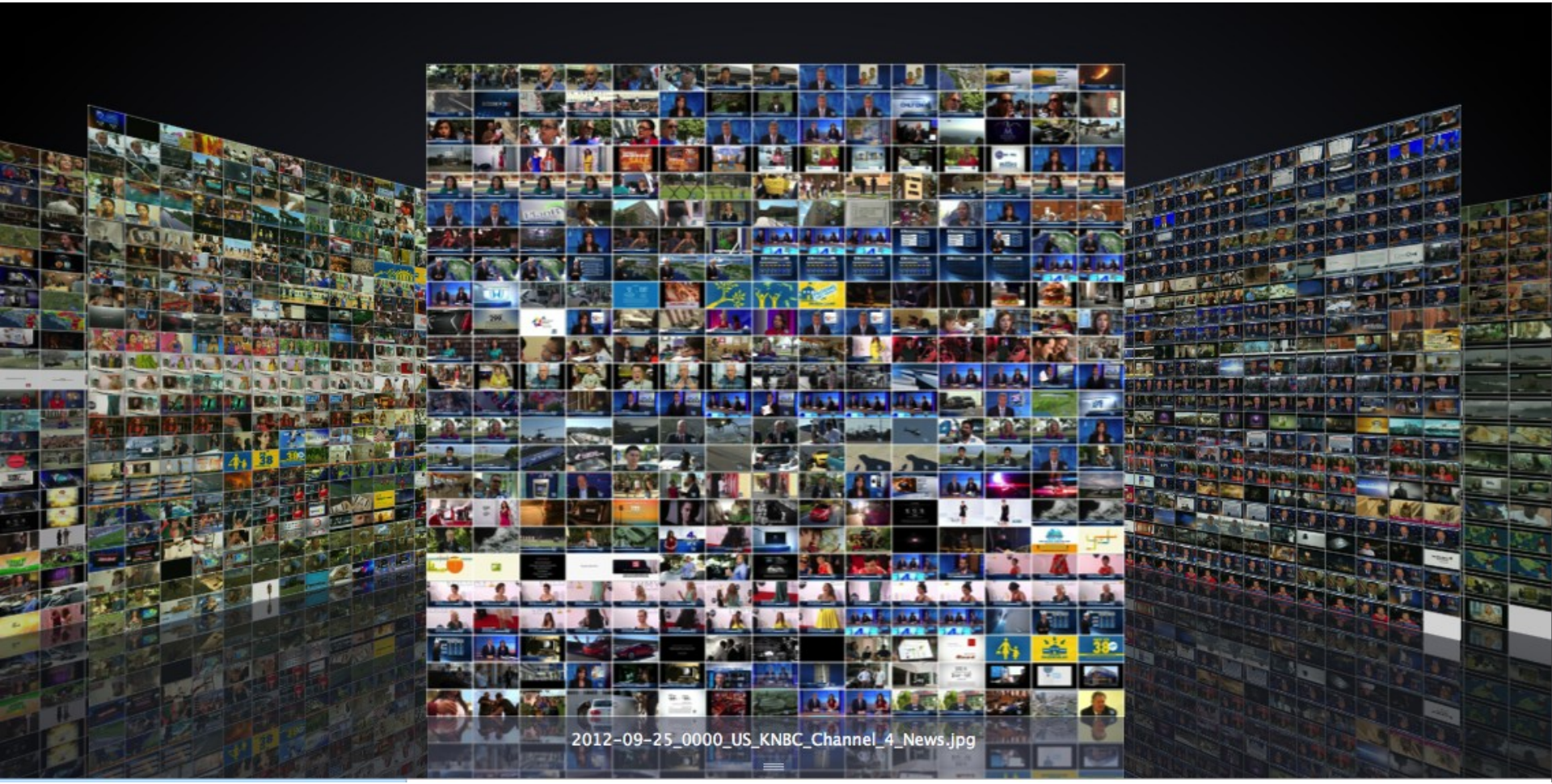
Natural Language Processing

- **Unstructured** digital content accounts for **90%** of all information [White paper IDC 2014] …
- Usually in the form of **texts** (also audio, video, etc.) and documents in multiple **languages** ...
- Only appropriate **NLP tools** can access this wealth of knowledge …
- NLP among the **top** 10 strategic technology trends for 2019 according to Gartner
- Spanish Plan for Language Technology 2015-2020

What happens in Internet every **second**? (july 2015)



What happens in Internet every **second**? (july 2015)



https://sites.google.com/site/distributedlittleredhen

… not only from Social Media

- LexisNexis receives **daily** 1.5M news.
- ?M judicial sentences, transcriptions ...
- ?M Electronic Health Records (EHR) …
- ?M Patents …
- … all kinds of e-documents …

- … and only **appropriate techniques** can handle all this digital data

Current AI challenges

- Natural Language Understanding
- Image/video Understanding
- Process/agents/services Understanding
- Database Understanding
- Web Understanding
- ...

Applying for a job at IKEA

Motivation

Sbdm ip im vdu yonrckblms.
Abf ip im vdu bhhigu.
Sbdm yigaus ly vdu hbbvfnoo.
Abf zumv vb vdu aivgdum.
Mduku ip vdu hbbvfnoo? A:yonrckblms
Mduku znp Abf fuhbku vdu aivgdum? A:bhhigu

Motivation

John is in the playground.
Bob is in the office.
John picked up the football.
Bob went to the kitchen.
Where is the football? A:playground
Where was Bob before the kitchen? A:office

Motivation

党

Motivation

מִפְלָגָה

Motivation

พรรค

Motivation

party

Motivation

party

- Which sense of “party”?
- How many senses have “party”?
- How these senses are translated to other languages?
- How a computer should represent these senses?
- How these senses combine to form phrases?

Motivation

The lexical-semantic knowledge allows us to better **characterize** the different **meanings** of the words

- *In 1992 Perot tried to organize a third **party** at the national level*
- *She joined the **party** after dinner*
- *They organized a **party** to search for food*
- *He planned a **party** to celebrate Bastille Day*
- *The **party** of the first part*

Motivation

- This better ***characterization*** may consist of:
- **Domain** tags to each word sense
  - party$^{1}_{n}$: politics
  - party$^{4}_{n}$: free-time
- **Semantic relations** that apply to each concept
  - party$^{1}_{n}$: member of: political_system$^{1}_{n}$
  - party$^{4}_{n}$: hyponym: wedding$^{1}_{n}$
- Lexical Knowledge Bases (LKBs)?
- Ontologies?
- Distributional Representations, now word embeddings?

Some personal background

- 1989-1992 Project **ACQUILEX**
- 1992-1995 Project **ACQUILEX II**
- 1996-1999 Project **EuroWordNet**
- 1998 **PhD** in Artificial Intelligence @ UPC
  - “Automatic Acquisition of Lexical Knowledge from MRDs”
  - DGILE Vox – Biblograf, some other Bilingual Dictionaries and WordNet
- 2002-2004 Project **MEANING**
  - Multilingual Central Repository (**MCR**)
- 2008-2011 **KYOTO**
- 2013-2015 **NewsReader**
- ...

Lexical Knowledge

- Which Knowledge is **needed** by a concrete NLP system?
- Where is this Knowledge **located**?
- Which automatic **procedures** can be applied?

Lexical Knowledge

- Which **Knowledge** is needed by a concrete NLP system?
  - Phonology: phonemes, stress, etc.
  - Morphology: lemma, POS, etc.
  - Syntactic: category, subcat., etc.
  - Semantic: class, Selectional Preferences, etc.
  - Pragmatic: usage, registers, domains, etc.
  - Translations: translation equivalences
  - ...

Lexical Knowledge

- Where is this Knowledge located?
  - Human **brain**
  - **Structured** Lexical Resources:
    - Monolingual and bilingual MRDs
    - Thesauri, encyclopedias
  - **Unstructured** Lexical Resources:
    - Monolingual and bilingual Corpora
  - Mixing resources

Lexical Knowledge

- Which automatic procedures can be applied?
  - **Prescriptive** approach (~ supervised)
    - Machine-aided manual construction
  - **Descriptive** approach (~ unsupervised)
    - Automatic acquisition from pre-existing Lexical Resources
  - Mixed approach

Where is this Knowledge located?

- Human brain:
  - **WordNet** (Miller et al. 90, Fellbaum 98)
    - Semantic network with >100,000 concepts
  - **CYC** ontology (Lenat 95, Malesh et al. 96, Matustek et al. 06)
    - 900 person-year of effort to produce 100,000 terms
  - **SUMO** (Niles & Pease 01, Niles & Pease 03)
    - IEEE ontology with ~25,000 terms and ~80,000 axioms
  - **VerbNet** (Kipper-Schuler 06), **FrameNet** (Baker et al. 98, Fillmore 12)
    - Verb lexicons with syntactic and semantic patterns
  - **UMLS** (Bodenreider 04)
    - Medical lexicon integrating 154 terminological resources for 25 languages

Where is this Knowledge located?

- Structured resources: monolingual MRDs
  - **LDOCE** Longman Dictionary for Contemporary English
    - learner's dictionary
    - 35,956 entries and 76,059 definitions
    - 86% semantic and 44% pragmatic codes
    - controlled vocabulary of 2,000 words
    - (Boguraev & Briscoe 89)
    - (Vossen & Serail 90)
    - (Bruce & Guthrie 92), (Wilks et al. 93)
    - (Dolan et al. 93), (Richardson 97)
    - (Green et al. 04)

Where is this Knowledge located?

- Structured resources: other MRDs
  - **Monolingual** MRDs
    - Webster's (Jensen & Ravin 87)
    - LPPL (Artola 93)
    - DGILE (Castellón 93), (Taulé 95), (Rigau 98), (Climent 98)
    - CIDE (Harley & Glennon 97)
    - AHD (Richardson 97)
    - WordNet (Harabagiu 98)
  - **Bilingual** MRDs
    - Collins Spanish/English (Knigth & Luk 94)
    - Vox/Harrap's Spanish/English (Rigau 98)

Where is this Knowledge located?

- Structured resources: thesauri and encyclopaedia
  - **Roget's Thesaurus**
    - 60,071 words in 1,000 categories
    - (Yarowsky 92), (Grefenstette 93), (Resnik 95), (Jarmasz 12)
  - **Wikipedia**
    - ~ 48M content articles, 294 active wikipedias
    - A source for mining meaning (Medelyan et al 09)
    - Also used as multilingual corpora
    - **DBpedia** (Lehmann et al. 12)
    - **BabelNet** (Navigli and Ponzetto 12)
  - but also **Wiktionary, Wikidata, OmegaWiki ...**

Where is this Knowledge located?

- Unstructured resources: corpora
  - **Monolingual** and **Bilingual**
  - **Raw** or **annotated**
  - **Main** source of lexical knowledge
  - **Distributional Semantics** (Distributional hypothesis):
    - *Words that are used and occur in the same contexts tend to purport similar meanings* (Harris 54)
    - *a word is characterized by the company it keeps* (Firth 57)
    - Basis of modern **Statistical Semantics** and **Neural Language Models**

    - Sketch Engine (Kilgarriff et al. 14, Kunilovskaya 17)
    - NoSketchEngine (Rychlý 07)

Acquisition of Lexical Knowledge from MRDs

- Why **MRDs**?
  usually "*contain spelling, pronunciation, hyphenation, capitalization, usage notes for semantic domains, geographic regions, and propiety; etymological, syntactic and semantic information about the most basic units of the language*" (Amsler 81)

- MRDs describe the **world** in a particular **language**
- But:
  - Conventional dictionaries are **not systematic**
  - Dictionaries are built for **human** use
  - **Implicit** Knowledge
    - words are described/translated in terms of words

Acquisition of Lexical Knowledge from MRDs

flor Organo complejo de la reproducción sexual en las plantas fanerógamas, procedente de la evolución de las hojas de un brote, y formado por órganos generadores de uno o de dos sexos, llamados estambres y pistilos, rodeados o no por las piezas de una envoltura o periantio simple, llamadas tépalos, o doble, llamadas sépalos y pétalos.

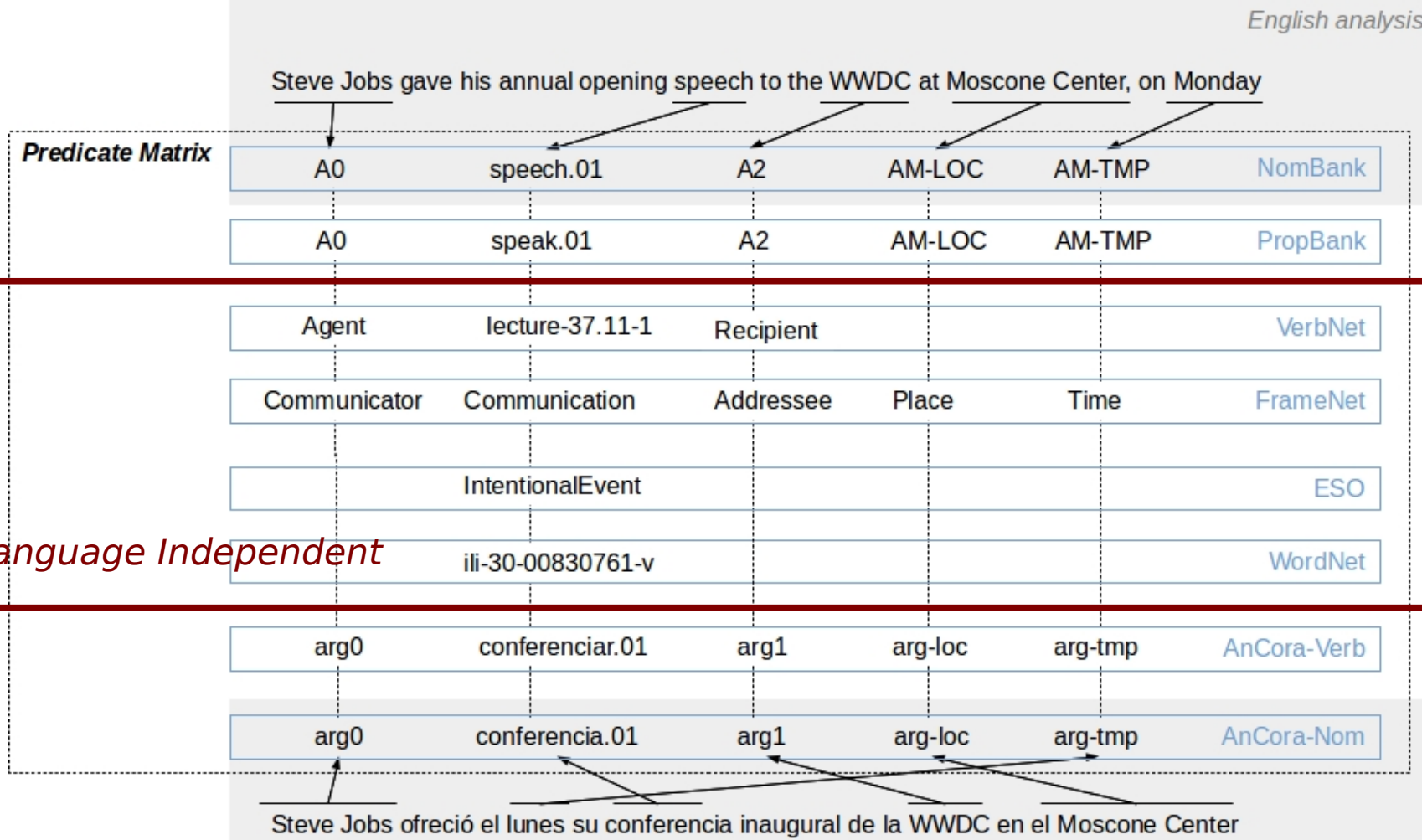
Acquisition of Lexical Knowledge from MRDs

jardín_1_1	Terreno donde se cultivan plantas y **flores** ornamentales.
florero_1_4	Maceta con **flores**.
ramo_1_3	Conjunto natural o artificial de **flores**, ramas o hierbas.
pétalo_1_1	Hoja que forma la corola de la **flor**.
tálamo_1_3	Receptáculo de la **flor**.
miel_1_1	Substancia viscosa y muy dulce que elaboran las abejas, en una distensión del esófago, con el jugo de las **flores** y luego depositan en las celdillas de sus panales.
florería_1_1	Floristería; tienda o puesto donde se venden **flores**.
florista_1_1	Persona que tiene por oficio hacer o vender **flores**.
camelia_1_1	Arbusto cameliáceo de jardín, originario de Oriente, de hojas perennes y lustrosas, y **flores** grandes, blancas, rojas o rosadas (Camellia japonica).
camelia_1_2	**Flor** de este arbusto.
rosa_1_1	**Flor** del rosal.

Global Wordnet Grid

- WordNet (Miller et al. 90, Fellbaum 98)
- Multilingual Central Repository (MCR) (Gonzalez-Agirre et al. 12)
  - EuroWordNet Framework: EN, ES, CAT, EUS, GAL, PO
  - WordNet Domains, BabelDomains, Base Concepts
  - Top Ontology, and the AdimenSUMO ontology
- Open Multilingual WordNet (Bond and Paik 12, Bond and Foster 13)
- GalNet (Gomez-Guinovart 11, Solla and Gomez-Guinovart 17)
- English WordNet (McRae et al. 19)

- Global WordNet Association
  - Wordnets in the world
  - Colaborative Inter-Lingual-Index (CILI)
  - Toward Multimodal WordNet workshop @ next LREC 2020


English analysis
Steve Jobs gave his annual opening speech to the WWDC at Moscone Center, on Monday
Predicate Matrix
A0 speech.01 A2 AM-LOC AM-TMP NomBank
A0 speak.01 A2 AM-LOC AM-TMP PropBank
Agent lecture-37.11-1 Recipient VerbNet
Communicator Communication Addressee Place Time FrameNet
IntentionalEvent ESO
ili-30-00830761-v WordNet
Language Independent
arg0 conferenciar.01 arg1 arg-loc arg-tmp AnCora-Verb
arg0 conferencia.01 arg1 arg-loc arg-tmp AnCora-Nom
Steve Jobs ofreció el lunes su conferencia inaugural de la WWDC en el Moscone Center
Spanish analysis

DEEP LEARNING - A NEW COMPUTING MODEL

"Training"

ImageNet

LEARNING ALGORITHM

"millions of trillions of FLOPS"

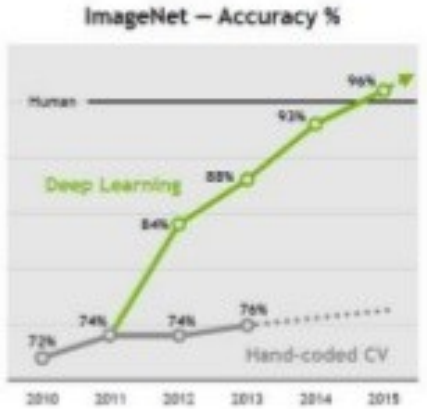
TB or PB of training data

Known Ground Truth Labels

"Inference"

"cat"

"SUPERHUMAN" RESULTS SPARK HYPERSCALE ADOPTION



From Andy Steinbach (NVIDIA)

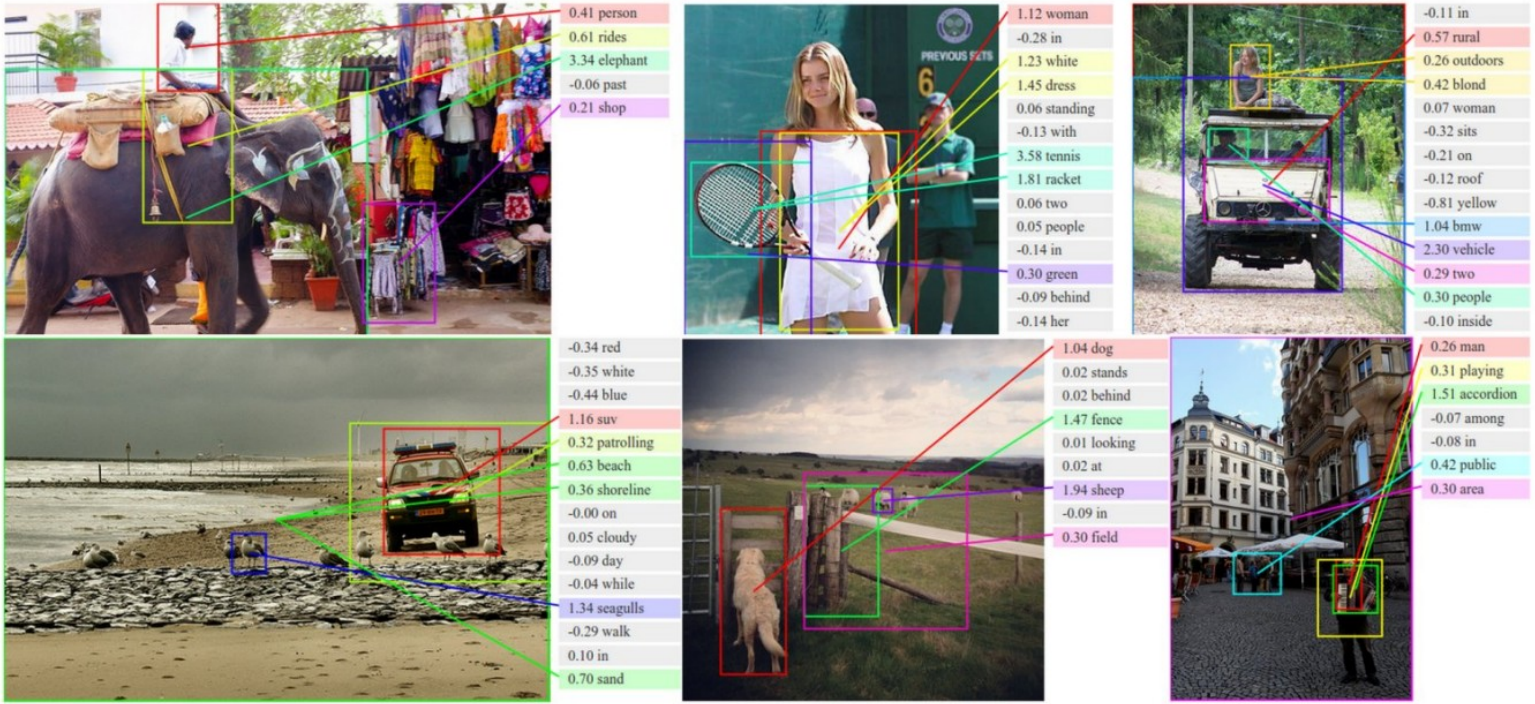


Figure 5. Example alignments predicted by our model. For every test image above, we retrieve the most compatible test sentence and visualize the highest-scoring region for each word (before MRF smoothing described in Section 3.1.4) and the associated scores ($v_i^T s_t$). We hide the alignments of low-scoring words to reduce clutter. We assign each region an arbitrary color.

Deep visual-semantic alignments for generating image descriptions (2014)

A Karpathy, L Fei-Fei

Figure 1. Photo-realistic images generated by our StackGAN from unseen text descriptions. Descriptions for birds and flowers are from CUB [32] and Oxford-102 [18] datasets, respectively. (a) Given text descriptions, Stage-I of StackGAN sketches rough shapes and basic colors of objects, yielding low resolution images. (b) Stage-II of StackGAN takes Stage-I results and text descriptions as inputs, and generates high resolution images with photo-realistic details.

StackGAN: Text to Photo-realistic Image Synthesis with Stacked Generative Adversarial Networks (2016)

Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaolei Huang, Xiaogang Wang, Dimitris Metaxas

LipNet: Sentence Level Lipreading (2016)

Yannis M. Assael, Brendan Shillingford, Shimon Whiteson, Nando de Freitas

Figure 2. Uncurated set of images produced by our style-based generator (config F) with the FFHQ dataset. Here we used a variation of the truncation trick [40, 5, 32] with $\psi = 0.7$ for resolutions $4^2 - 32^2$. Please see the accompanying video for more results.

A Style-Based Generator Architecture for Generative Adversarial Networks (2018)

Tero Karras, Samuli Laine, Timo Aila

https://thispersondoesnotexist.com

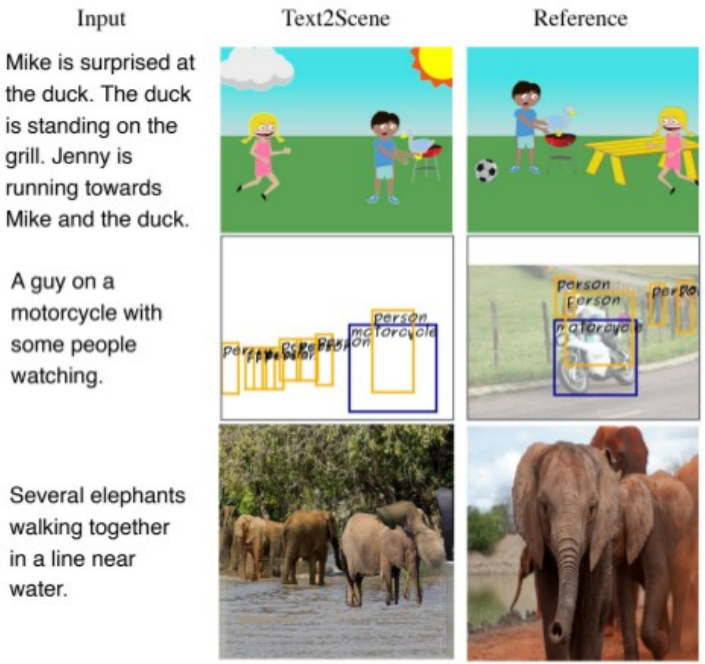


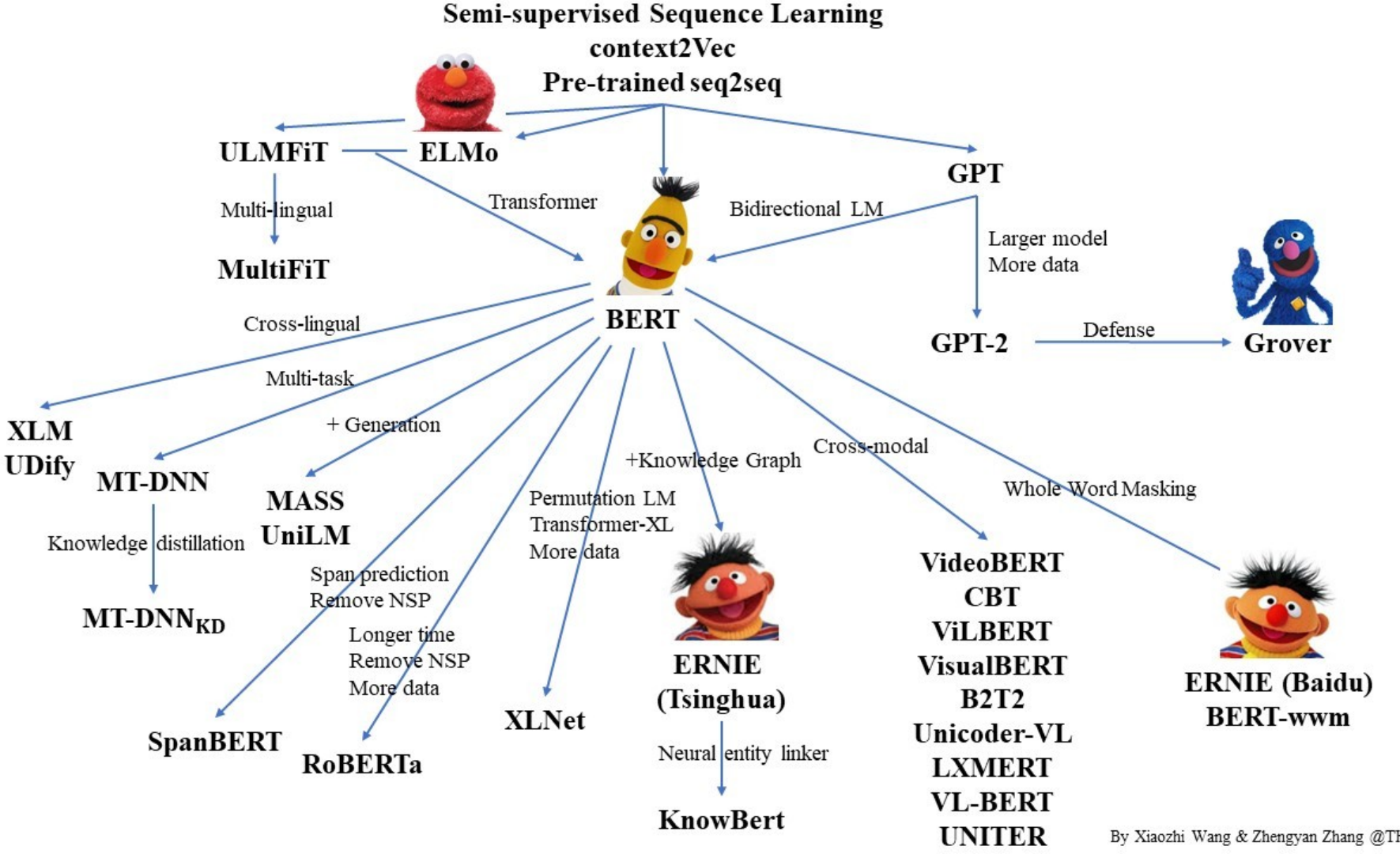
Figure 1. Sample inputs (left) and outputs of our Text2Scene model (middle), along with *ground truth* reference scenes (right) for the generation of abstract scenes (top), object layouts (middle), and synthetic image composites (bottom).

Text2Scene: Generating Compositional Scenes from Textual Descriptions (2018)

Fuwen Tan, Song Feng, Vicente Ordonez

Neural Language Models

- **Static** word embeddings
  - word2vec (Mikolov et al. 13a, 13b, 13c)
  - GloVe (Pennington et al. 14)
  - FastText (Bojanowski et al. 17)
  - Dict2vec (Tissier et al. 17)
- **Contextual** word embeddings
  - ELMo (Gardner et al. 17)
  - GPT-2 (Radford et al. 19)
  - BERT (Deblin et al. 19)
  - Transformers Talk2transformer
  - Glue leaderboard
- something2vec, awesome2vec …

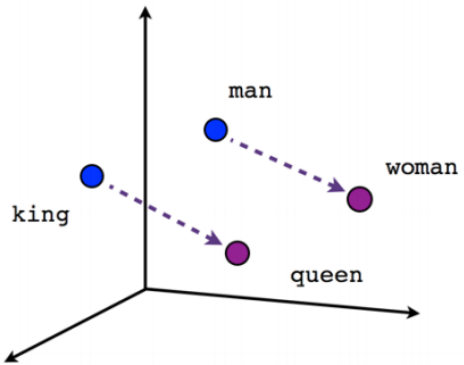

Semi-supervised Sequence Learning
context2Vec
Pre-trained seq2seq
ULMFiT
ELMo
GPT
Multi-lingual
MultiFiT
Transformer
Bidirectional LM
Larger model
More data
BERT
GPT-2
Defense
Grover
Cross-lingual
XLM
UDify
Multi-task
MT-DNN
Knowledge distillation
MT-DNN$_{KD}$
+ Generation
MASS
UniLM
Span prediction
Remove NSP
SpanBERT
Longer time
Remove NSP
More data
RoBERTa
Permutation LM
Transformer-XL
More data
XLNet
+Knowledge Graph
ERNIE
(Tsinghua)
Neural entity linker
KnowBert
Cross-modal
VideoBERT
CBT
ViLBERT
VisualBERT
B2T2
Unicoder-VL
LXMERT
VL-BERT
UNITER
Whole Word Masking
ERNIE (Baidu)
BERT-wwm
By Xiaozhi Wang & Zhengyan Zhang @THUNLP

word2vec

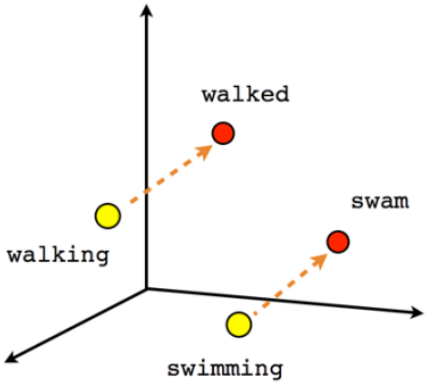
- **Fast** and **simple** neural models
- **Large** collections of texts
- Two flavours:
  - continuous bag-of-words (**CBOW**): context -predict→ word
  - **Skip-gram**: word -predict→ context
- **Unsupervised**, negative sampling
- **Words** as vectors with a few hundreds of weights (**Vector Space**)
- **No** feature engineering

- *brother - boy + girl ~ sister*
- *queen - woman + man ~ king*
- *biking - today + yesterday ~ biked*

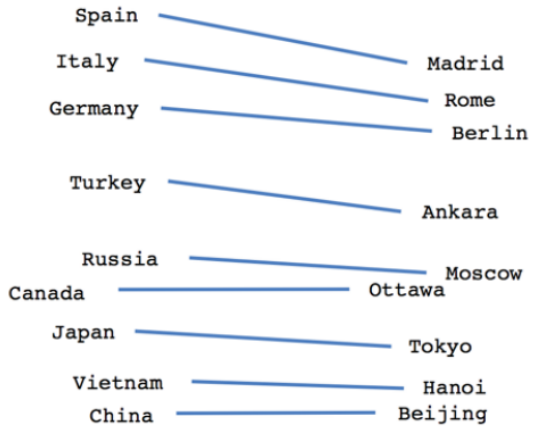
word2vec



Male-Female



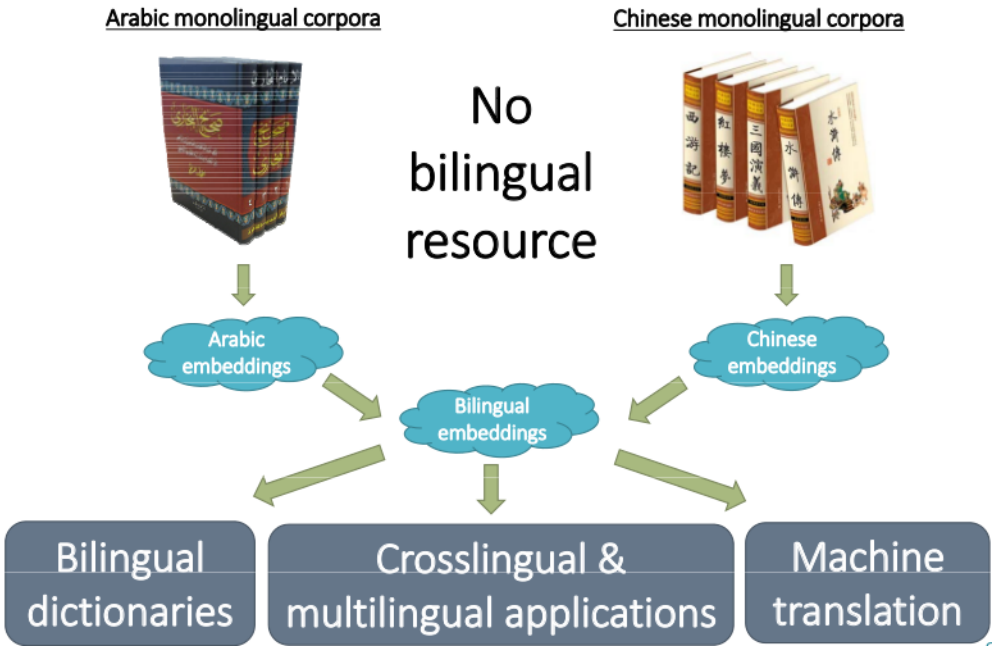
Verb tense



Country-Capital

- **Meaningful** distances and relations (!) … **similarity**, **analogy** …
- Languages are (to a large extent) **isometric** in word embedding space (!)

Cross-lingual Word Embeddings



https://github.com/facebookresearch/MUSE (Conneau et al. 17)
https://github.com/artetxem/vecmap (Artetxe et al. 16, 17, 18a, 18b)

… and back!

Word	Generated definition
brawler	a person who fights
butterfish	a marine fish of the atlantic coast
continually	in a constant manner
creek	a narrow stream of water
feminine	having the character of a woman
juvenility	the quality of being childish
mathematical	of or pertaining to the science of mathematics
negotiate	to make a contract or agreement
prance	to walk in a lofty manner
resent	to have a feeling of anger or dislike
similar	having the same qualities
valueless	not useful

Table 1: Selected examples of generated definitions. The model has been trained on occurrences of each example word in running text, but not on the definitions.

https://github.com/websail-nu/torch-defseq (Noraset et al. 2017)

… and back!

Word	Context	Definition
star	she got star treatment	a person who is very important
star	bright star in the sky	a small circle of a celestial object or planet that is seen in a circle
sentence	sentence in prison	an act of restraining someone or something
sentence	write up the sentence	a piece of text written to be printed
head	the head of a man	the upper part of a human body
head	he will be the head of the office	the chief part of an organization, institution, etc
reprint	they never reprinted the famous treatise	a written or printed version of a book or other publication
rape	the woman was raped on her way home at night	the act of killing
invisible	he pushed the string through an inconspicuous hole	not able to be seen
shake	my faith has been shaken	cause to be unable to think clearly
nickname	the nickname for the u.s. constitution is 'old ironsides '	a name for a person or thing that is not genuine

Table 2: Examples of definitions generated by *S + I-Attention* model for the words and contexts from the test set.

https://github.com/agadetsky/pytorch-definitions (Gadetsky et al. 2018)

Semantics vs pragmatics

Questions from lawyers that were taken from official court records (I)

- Q: You were there until the time you left, is that true?
- Q: You don't know what it was, and you didn't know what it looked like, but can you describe it?
- Q: How far apart were the vehicles at the time of the collision?
- Q: Was it you or your brother that was killed in the war?
- Q: How many times have you committed suicide?

Semantics vs pragmatics

Questions from lawyers that were taken from official court records.(II)

- Q: What happened then?
  A: He told me 'I have to kill you because you can identify me.'
  Q: Did he kill you?
- Q: Do you recognize that picture?
  A: That's me.
  Q: Were you present when that picture was taken?
- Q: Now, Mrs. Johnson, how was your first marriage terminated?
  A: By death.
  Q: And by whose death was it terminated?

Semantics vs pragmatics

Questions from lawyers that were taken from official court records.(III)

- Q: "Doctor, before you performed the autopsy, did you checked its pulse?"
- A: "No"
- Q: "Did you check its blood pressure?"
- A: "No"
- Q: "Have you checked if he was breathing?"
- A: "No"
- Q: "So, it is possible that the patient was alive when you began the autopsy?"
- A: "No"
- Q: "How can you be so sure, Doctor?"
- A: "Because his brain was on my desk in a jar"
- Q: "But could, however, the patient have still been alive?"
- A: "It may have been alive and practicing law somewhere."

http://ixa2.si.ehu.es/deep_learning_seminar/

Deep Learning neural network models have been successfully applied to natural language processing, and are now changing radically how we interact with machines (Siri, Amazon Alexa, Google Home, Skype translator, Google Translate, or the Google search engine). These models are able to infer a continuous representation for words and sentences, instead of using hand-engineered features as in other machine learning approaches. The seminar will introduce the main deep learning models used in natural language processing, allowing the attendees to gain hands-on understanding and implementation of them in Tensorflow.

ixa2.si.ehu.eus/master - Mozilla Firefox

Inbox (21,525) - grigau@ | TAIA 2019-2020 - Google | Artificial Intelligence - Go | Google Calendar - Week | ixa2.si.ehu.eus/master | Review Request Accepte

https://ixa.eus/master/index.php?lang=en

EU · EN

HOME · HAP-LAP · EMLCT (2017-2019) · IXA GROUP · CONTACT

Hizkuntzaren Azterketa eta Prozesamendua
Language Analysis and Processing

(C) Joserra Martínez (Txerra) eta Sacha Tovar

Introduction

https://ixa.eus/master

Advanced applications, based on the use of Language Analysis and Processing techniques, such as machine translation, automatic speech recognition and synthesis, dialogue systems, question answering, information extraction and web search systems are fast becoming an integrated part of our digital devices and daily lives. They are a key component in artificial intelligence-based solutions that provide people with natural ways to communicate with machines (e.g. Siri or Alexa), other people (e.g. machine translation in Skype), information repositories and much more.

These technologies require a mix of skills and approaches from computer science, linguistics, statistics, artificial intelligence, machine learning, logic... If you are a computer specialist, an engineer or a mathematician, you may be wondering about the basic computing tools working behind these devices or how you can integrate these applications within other applications or the web. If you are a philologist or linguist, you probably wonder how linguistic knowledge is used in them: lexicon, grammar and word-senses; morphological, syntactic and semantic analyses...




http://hitz.eus/

HiTZ: Basque Center for Language Technology

HiTZ is a reference center on Language Technologies. Its aim is to promote research, training, technological transfer and innovation in this area. We are a multidisciplinary team: computer scientists, linguists and engineers.

- **Information Extraction and Information Retrieval**
- **Machine Translation**
- **Text Analysis**
- **Speech Synthesis and Speech Recognition**

Computational Lexicography: from dictionaries to deep neural networks, and back

German Rigau <german.rigau@ehu.eus>