Using deep learning paradigm to improve syllabic versification: A first approach

Javier de la Rosa (versae@linhd.uned.es, @versae)
Postdoctoral Fellow ERC Poetry Standardization and Linked Open Data (POSTDATA) Project
Salvador Ros(sros@ssc.uned.es)
Technical Director and Co-PI ERC Poetry Standardization and Linked

Laboratorio de Innovación en Humanidades Digitales, UNED, Spain

European Research Council

THE FRAMEWORK PROGRAMME FOR RESEARCH AND INNOVATION
HORIZON 2020

BERTSIFICATION
Uncovering metrical form hidden in words embeddings

Javier de la Rosa (versae@linhd.uned.es, @versae)
Postdoctoral Fellow ERC Poetry Standardization and Linked Open Data (POSTDATA) Project
Salvador Ros(sros@ssc.uned.es)
Technical Director and Co-PI ERC Poetry Standardization and Linked

Laboratorio de Innovación en Humanidades Digitales, UNED, Spain

UNED

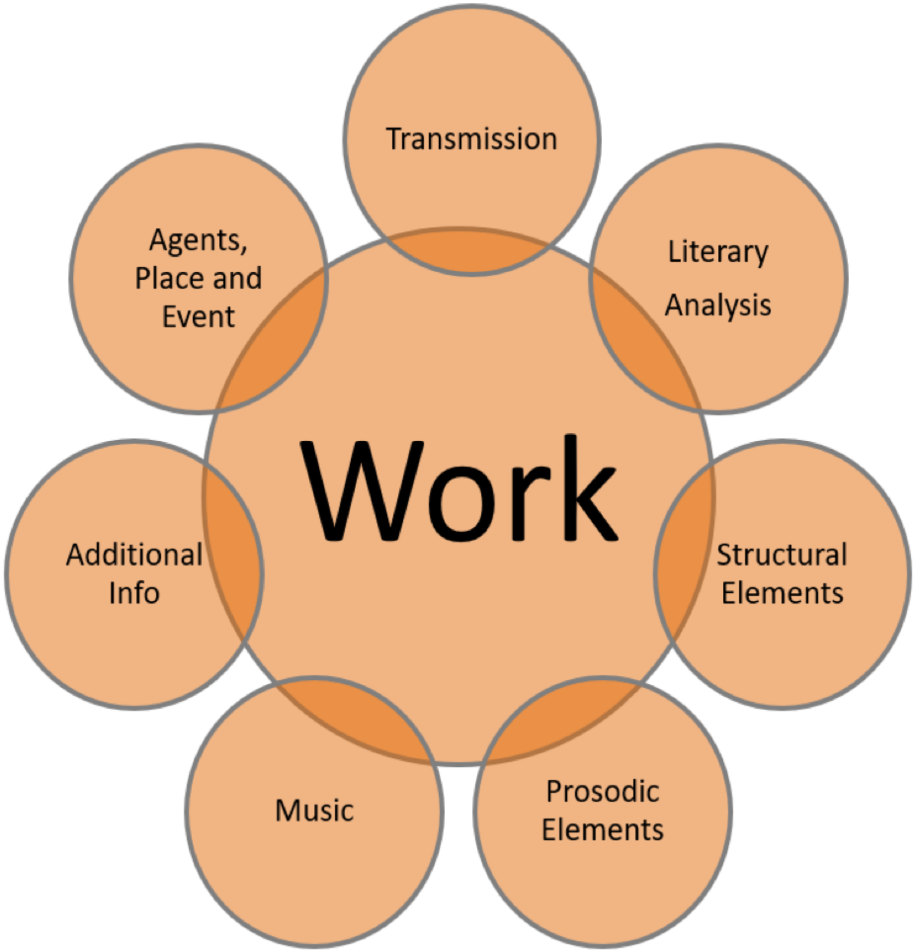
LiNHD
LABORATORiO DE iNNOVACiÓN EN HUMANiDADES DiGiTALES

POSTDATA
Poetry Standardization and Linked Open Data

POSTDATA

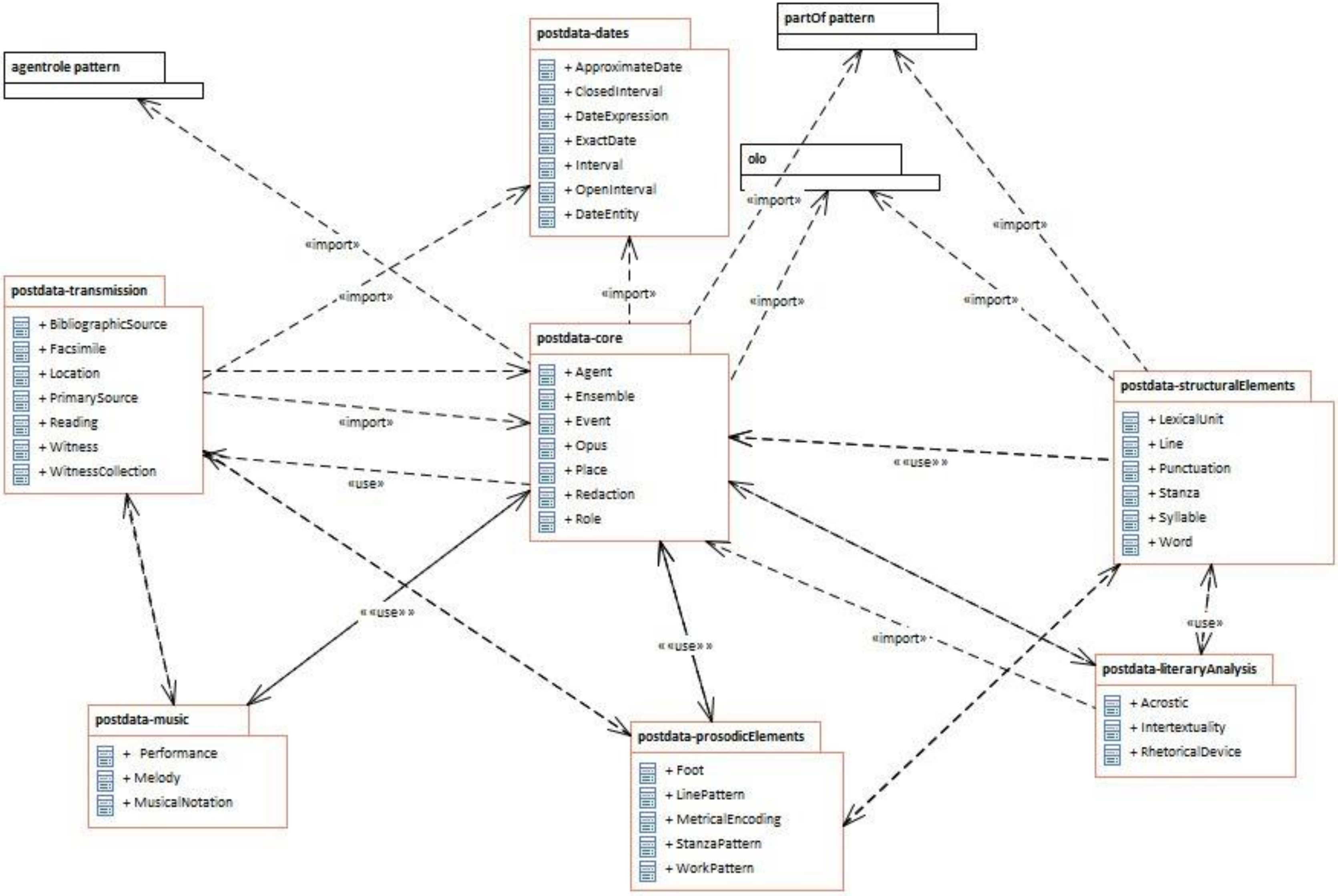
- ERC Starting Grant
- Poetry Standardization and Linked Open Data

POSTDATA


Transmission
Agents, Place and Event
Literary Analysis
Work
Additional Info
Structural Elements
Music
Prosodic Elements

POSTDATA

POSTDATA
Poetry Standardization
and Linked Open Data



agentrole pattern

postdata-dates
+ ApproximateDate
+ ClosedInterval
+ DateExpression
+ ExactDate
+ Interval
+ OpenInterval
+ DateEntity

partOf pattern

olo

postdata-transmission
+ BibliographicSource
+ Facsimile
+ Location
+ PrimarySource
+ Reading
+ Witness
+ WitnessCollection

postdata-core
+ Agent
+ Ensemble
+ Event
+ Opus
+ Place
+ Redaction
+ Role

postdata-structuralElements
+ LexicalUnit
+ Line
+ Punctuation
+ Stanza
+ Syllable
+ Word

postdata-literaryAnalysis
+ Acrostic
+ Intertextuality
+ RhetoricalDevice

postdata-music
+ Performance
+ Melody
+ MusicalNotation

postdata-prosodicElements
+ Foot
+ LinePattern
+ MetricalEncoding
+ StanzaPattern
+ WorkPattern

POSTDATA

- Very rich model → Hard to fill
- How can we help in filling it?

POSTDATA

- Very rich model → Hard to fill

- How can we help in filling it?
  - Automated scansion

Automated scansion

- Rule-based

- Inference-based
  - Generative
  - Neural scansion
  - Machine learning

Rule-based scansion

- Syllabification

- Prosody and stress

- Patterns and exceptions

Rule-based scansion

«Vino y ahogó sus penas»

- Syllabification
  - Vi-no y a-ho-gó sus pe-nas
- Prosody and stress
  - **Vi**-no‿**y**‿a‿ho-**gó**-sus-**pe**-nas
- Patterns and exceptions
  - **Vi**-no-y‿a‿ho-**gó**-sus-**pe**-nas

Rule-based scansion

- English

- Spanish

Rule-based scansion for English

- Scandroid
- Zeuscansion
- Poesy (aka LitLab-poetry)

- Others: Calliope, AnalysePoems, etc.

Rule-based scansion for English

- Scandroid [1996, 2005]
  - iambic or anapestic meter
- Zeuscansion [2016]
  - OOV and stress-guesser
- Poesy (aka LitLab-poetry) [2018]
  - Diificult coding of new rhythmic patterns

Rule-based scansion for Spanish

- Gervás [2000]
- ADSO Scansion System [2017]
- SKAS [2017]

Rule-based scansion for Spanish

- Gervás [2000]
  - Logic programming
- ADSO Scansion System [2017]
  - Special focus on synaloephas
- SKAS [2017] → PoetryLab Ran·tan·plan
  - Industrial strong NLP
  - SpaCy models and API

Neural scansion

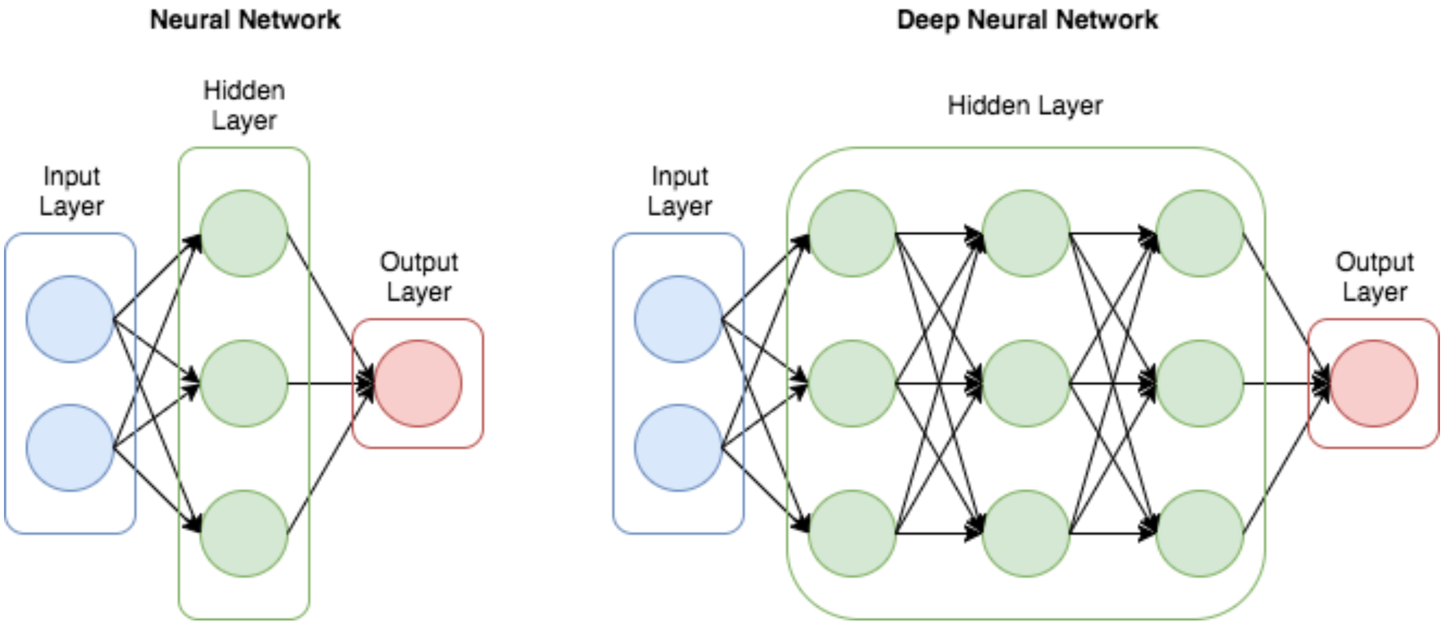
- IXA (Spain) [2018]
  - English and Spanish

- KAIST (Korea) [2019]
  - English

Neural scansion

POSTDATA
Poetry Standardization
and Linked Open Data

- Same approach BiLSTM-{CNN,CRF}

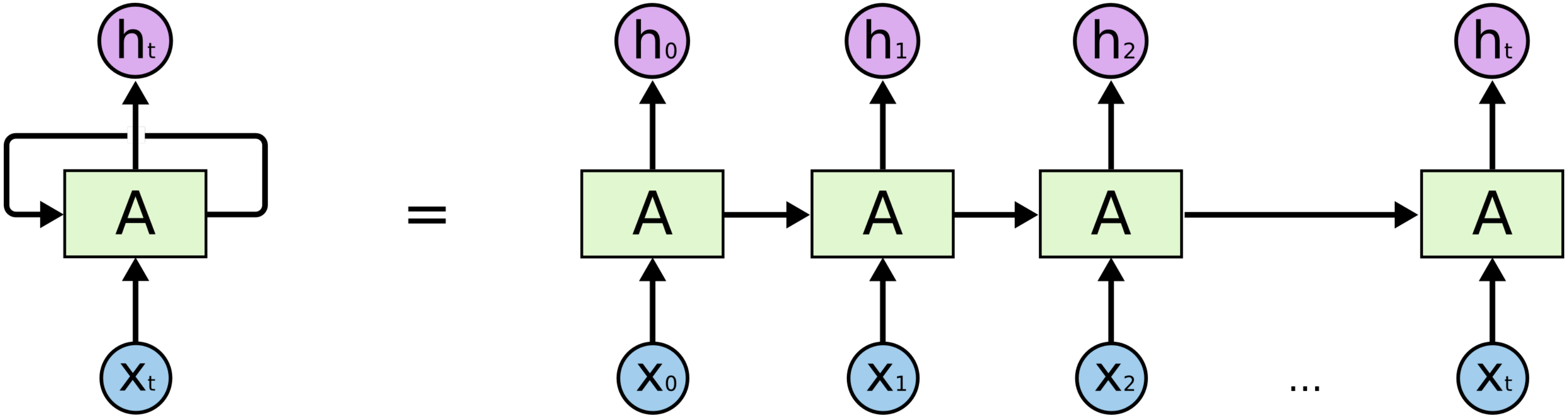


Neural Network

Source: https://www.reddit.com/r/learnmachinelearning/comments/9m561w/how_do_i_accurately_understand_a_neural_network/

Neural scansion

- Same approach BiLSTM-{CNN,CRF}

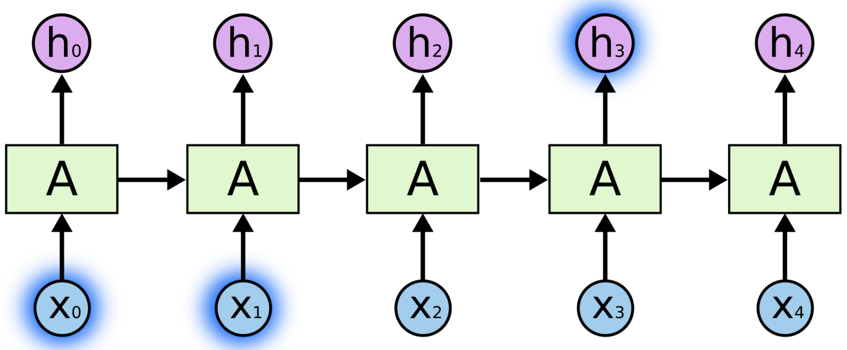


RNN's

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

Neural scansion

- Same approach BiLSTM-{CNN,CRF}

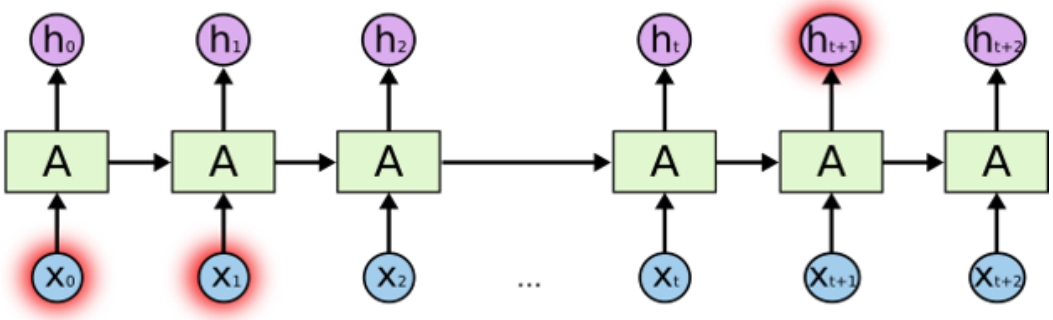


"the clouds are in the *sky*"

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

Neural scansion

- Same approach BiLSTM-{CNN,CRF}



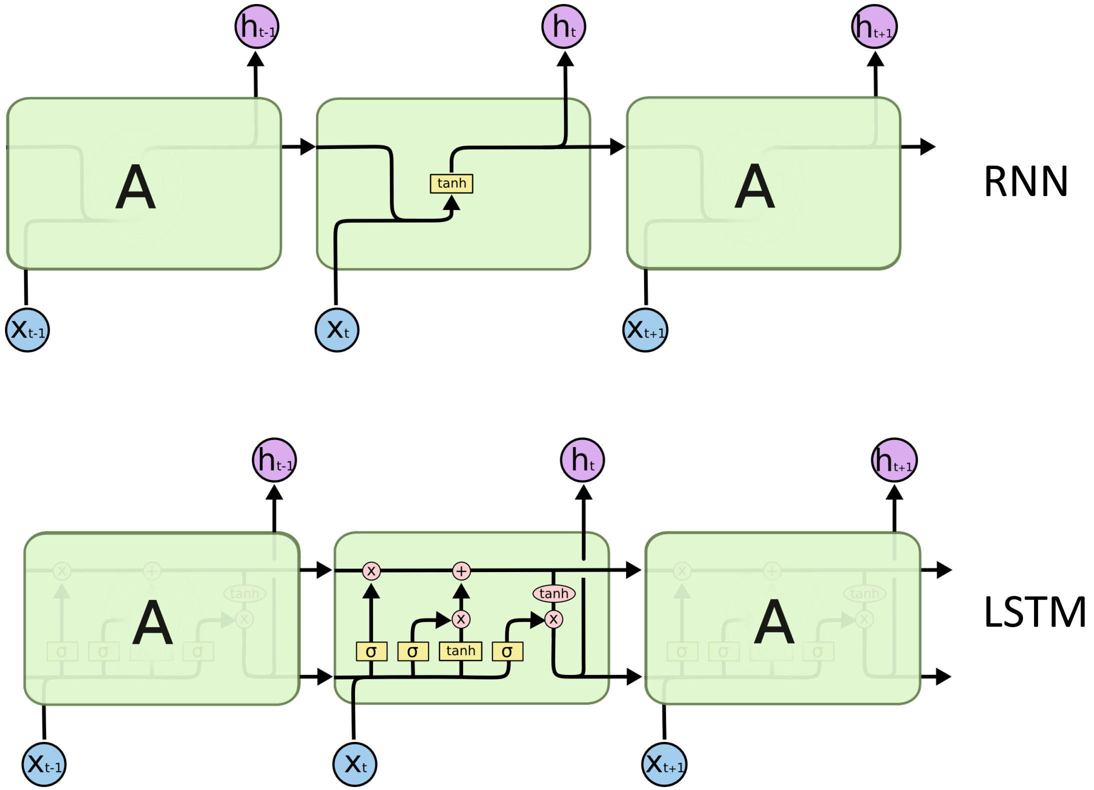
"I grew up in France… I speak fluent *French*."

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

Neural scansion

POSTDATA
Poetry Standardization
and Linked Open Data

- Same approach BiLSTM-{CNN,CRF}

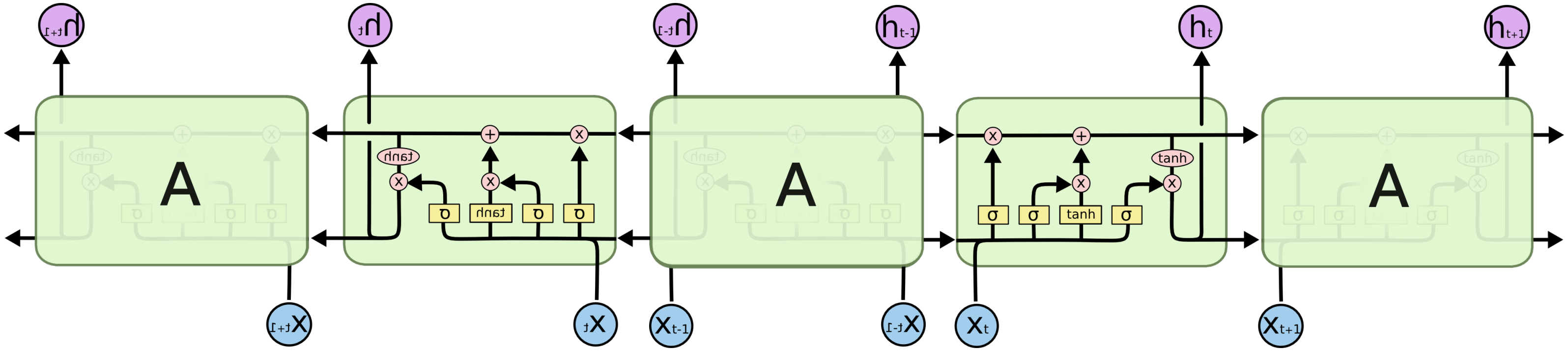


Source: http://colah.github.io/posts/2015-08-Understanding-LSTMs/

Neural scansion

POSTDATA
Poetry Standardization
and Linked Open Data

- Same approach BiLSTM-{CNN,CRF}

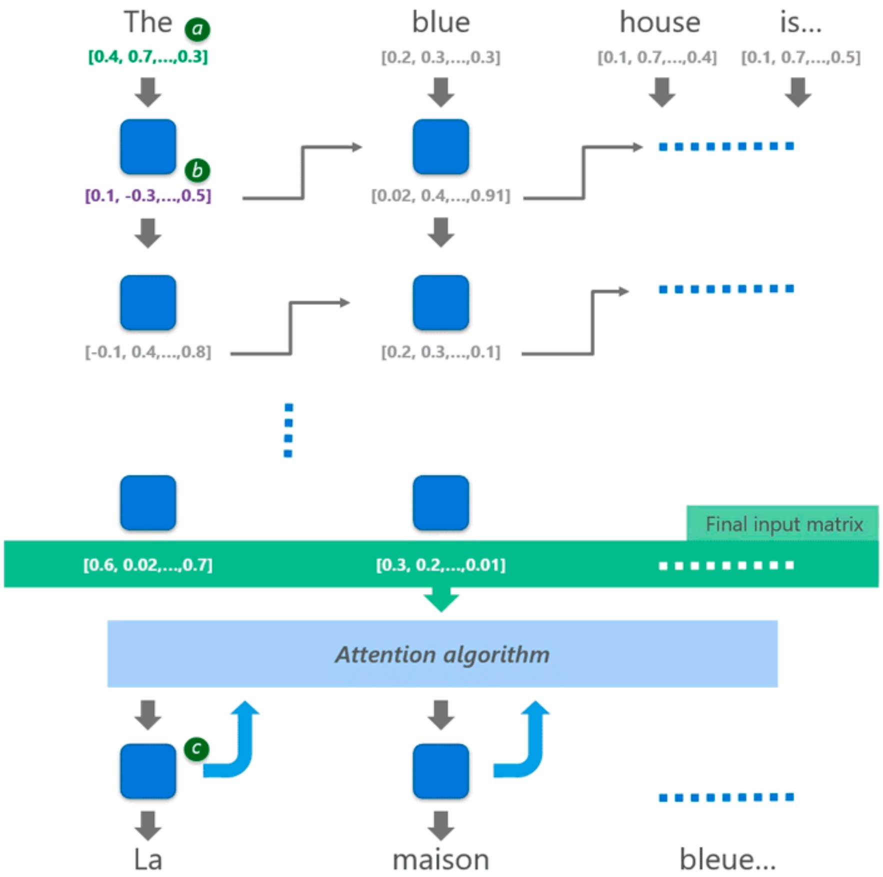


Bidirectional LSTM

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

Neural scansion

- Attention mechanism (counts + context)



Neural scansion

- BiLSTM with Attention
  - Our results
  - The problem of the corpus

Neural scansion

- BiLSTM with Attention
  - Our results → 92.43% per verse (90.84% SOTA)
  - The problem of the corpus
    - Borja Navarro's annotated corpus
    - Hendecasyllables with stress in penultimate position
    - Mixed manual and automated (rule-based) annotation

Neural scansion

- BiLSTM with Attention
  - Our results → 92.43% per verse (90.84% SOTA)
  - The problem of the corpus
    - Borja Navarro's annotated corpus
    - Hendecasyllables with stress in penultimate position
    - Mixed manual and automated (rule-based) annotation

END OF A FIRST APPROACH

Neural scansion

- Word embeddings

  “You shall know a word by the company it keeps”

  -- J.R. Firth

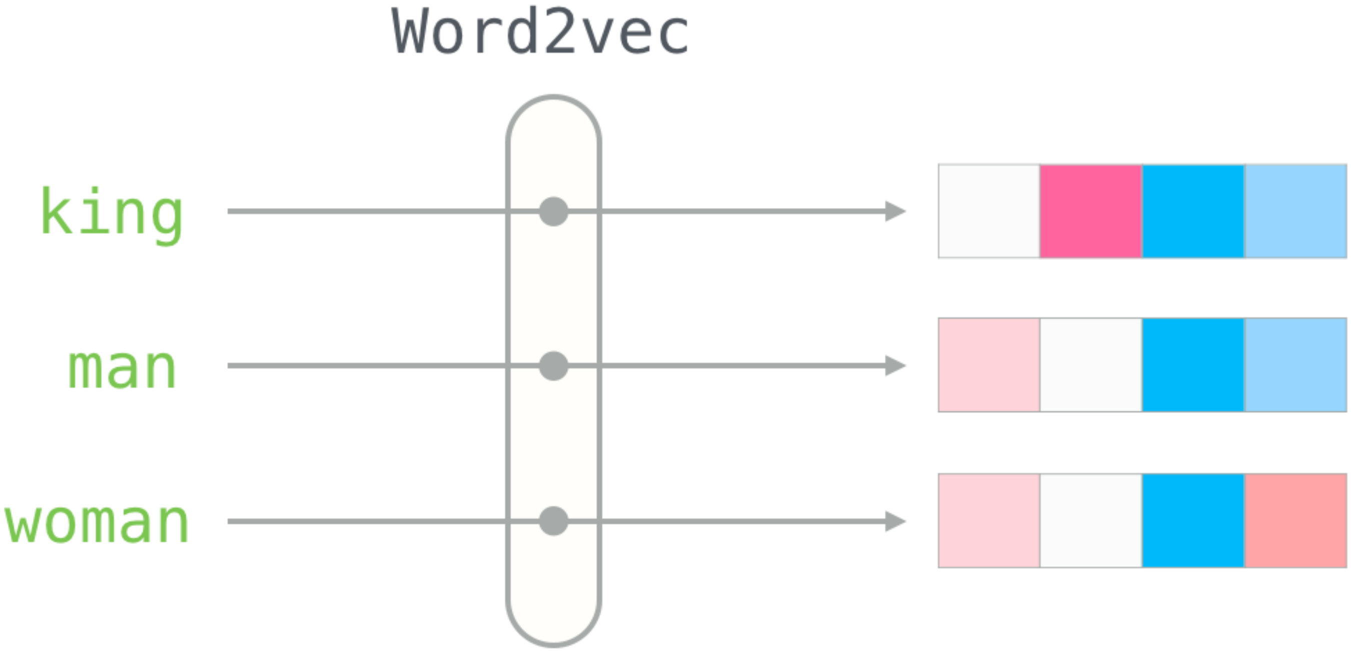
Neural scansion

- Other applications of Attention
  - Word and sentence embeddings
  - Contextualized word embeddings
    - GPT-2 (Transformer)
    - ULMFit
    - ELMO
    - BERT

Neural scansion

POSTDATA
Poetry Standardization and Linked Open Data

- Word embeddings

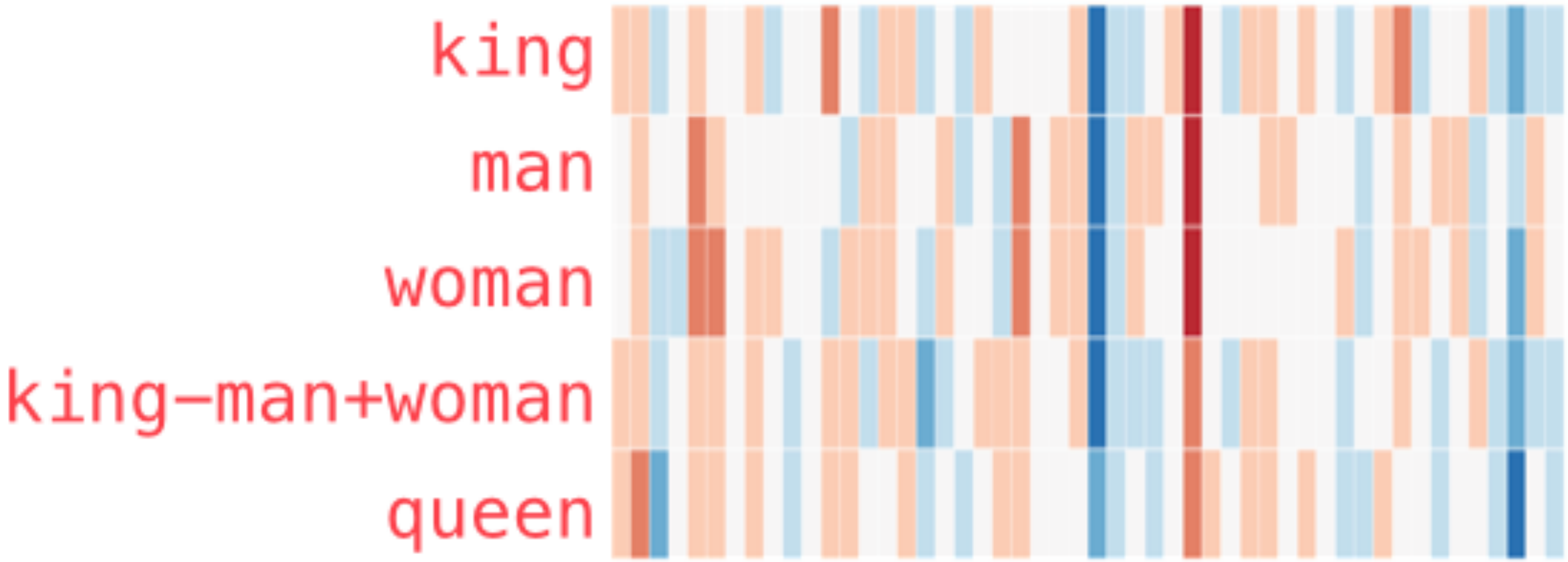


Source: http://jalammar.github.io/illustrated-word2vec/

Neural scansion

- Word embeddings

king – man + woman ~= queen



Source: http://jalammar.github.io/illustrated-word2vec/

Neural scansion

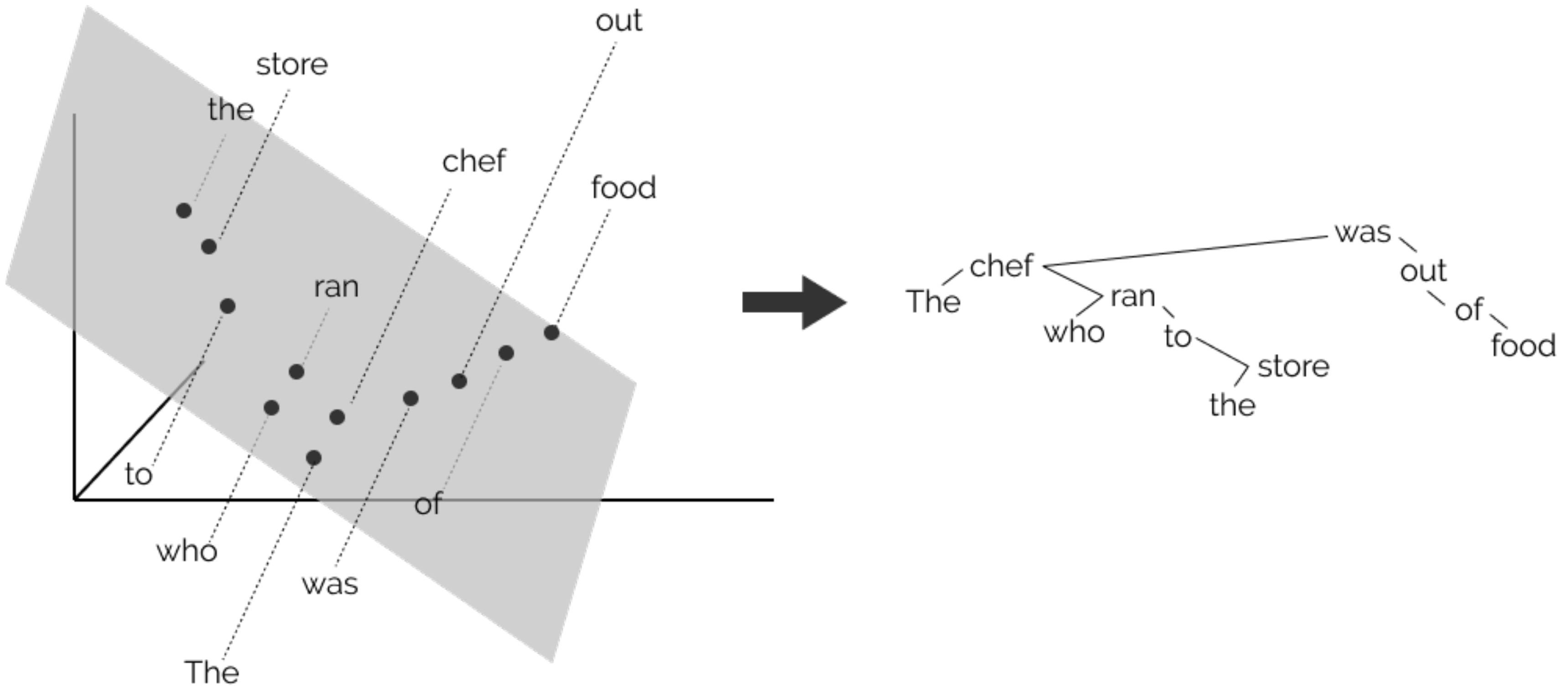
- Word embeddings

  "What you can cram into a single $&!#* vector: Probing sentence embeddings for linguistic properties" Alexis Conneau *et. Al.*

  – SentLen, WC, TreeDepth, TopConst, BShift, Tense, SubjNum, ObjNum, SOMO, CoordInv

Neural scansion

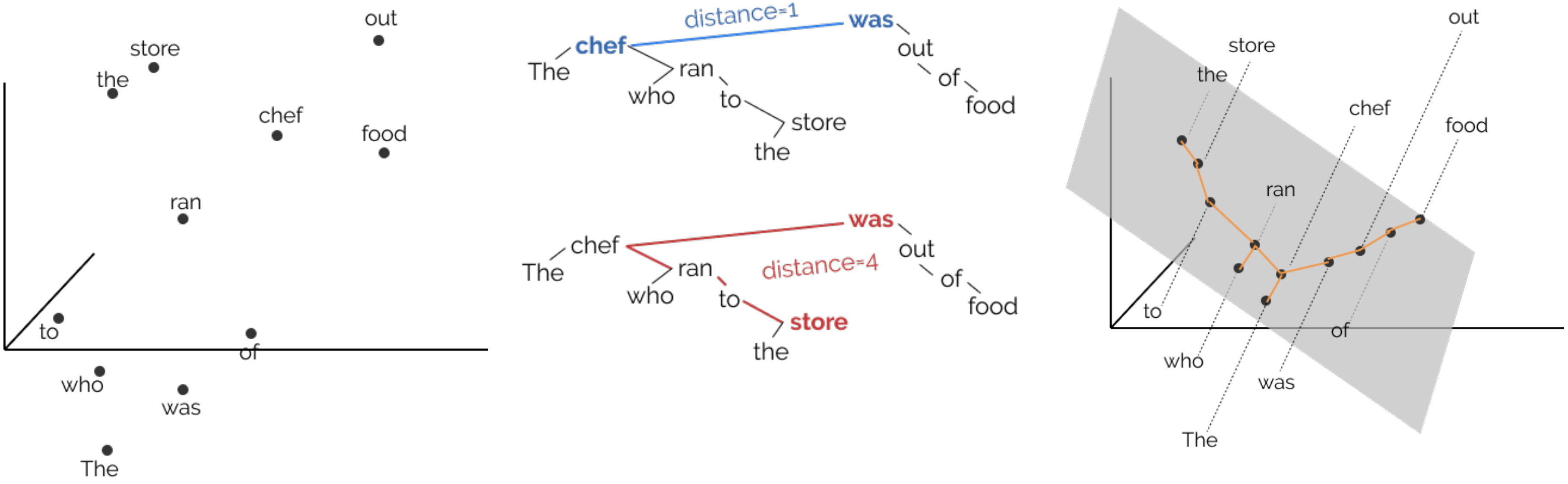
- Probing contextualized sentence embeddings



Source: https://nlp.stanford.edu/~johnhew/structural-probe.html

Neural scansion

- Probing contextualized sentence embeddings



Source: https://nlp.stanford.edu/~johnhew/structural-probe.html

Neural scansion

- Other applications of Attention
  - Word and sentence embeddings
  - Contextualized word embeddings
    - GPT-2 (Transformer)
    - ULMFit
    - ELMO
    - **BERT (Bidirectional Encoder Representations from Transformers)**

Neural scansion: our approach

- BERT
  - Multilingual and cased
  - Pre-trained model as features / Pre-train and fine-tuning
  - Successful probing

Neural scansion: our approach

- Probing BERT (Flair) for meter information
  - **Vi**-no-y‿a‿ho-**gó**-sus-**pe**-nas

+ - - + -

+ - 7

Neural scansion: our approach

- Advantages
  - Easily multi-lingual
  - Access to pre-trained models
  - No prior fixed length requirement
  - Decent **accuracy (grouped for now!)**

Neural scansion: our approach

- Decent **enough**?
  - What does this even mean?
  - For what purpose?
  - How uncertainty should be registered or catalogued?

Neural scansion: our approach

- Disadvantages
  - Complex construction
  - Need for pre-trained models built using massive computing infrastructures

Conclusions

- Contextualized sentence/verse embeddings might generalize well
- Paving the way for an unsupervised multi-language scansion system
- Need better corpora!

Questions and Thanks!

Javier de la Rosa
versae@linhd.uned.es
& LINHD Team

http://postdata.linhd.uned.es/
http://linhd.uned.es
@linhduned

LiNHD
LABORATORiO DE iNNOVACiÓN
EN HUMANiDADES DiGiTALES

POSTDATA
Poetry Standardization
and Linked Open Data