Learning neural Question Answering Systems for Low resource Langauges

Overview

1. Problem statement: what is question answering
2. Neural question answering and the challenges for resource-poor languages
3. Proposed approaches to transfer learning
4. Experiment design
5. Experiment Results
6. Conclusions

What’s Question Answering

- General Sense:
  - Information retrieval: web queries, smart assistants, etc.
  - Natural language understanding: Reading comprehension test
  - Combinations: Jeopardy! etc.
- Narrower Definition:
  - Document + Question + Answer → Model
  - Model (Document, Question) → Answer

What’s Question Answering

(@entity3) the @entity2 military claims to have shot down a @entity5 drone , state media reported tuesday . " @entity2 air defense systems shot down a @entity5 @entity9 (unmanned aerial vehicle) north of @entity11 , " the state - run @entity6 said . the @entity5 military lost contact with a mq - 1 predator drone over @entity2 , a @entity5 official said tuesday …

@placeholder 's military claims it shot down a @entity5 drone north of @entity11

(@entity18) – Un hombre fue detenido por su presunta participación en un tiroteo en el que murieron tres personas durante una reunión, en una pequeña ciudad del estado de @entity8, en la costa oeste de @entity5, dijeron las autoridades. En el tiroteo -registrado en @entity11, a unos 45 kilómetros al norte de @entity13- también resultó herida una persona. @entity0, …

El tiroteo se registró en el poblado de @placeholder

Teaching Machines to Read and Comprehend, Hermann et al., NIPS 2015

Neural QA Systems

- Memory Networks (Weston et al. 2015)
  - Effective on simple logical statements
- N2NMenNet, DMN, etc.
- Attentive Reader (Hermann et al. 2015)
  - News article reading comprehension
- Neural Turing Machines (Graves et al. 2014), Differentiable Neural Computers (Graves et al. 2016)
  - Various tasks (logic statements, graph reasoning, etc.)

The Research Question

- Large amount of data is needed to train a neural QA system
  - 380K Q&A pairs in the CNN news dataset (machine-generated)
  - 120K Q&A pairs in the Maluuba NewsQA dataset (human-annotated)
  - 100,000+ Q&A pairs in SQUAD dataset
  - Such large datasets (especially human-annotated ones) are difficult to compile
- But the data doesn’t exist for many languages
  - Even for relatively widely-spoken languages like French, Spanish or Chinese
- How to solve this problem?

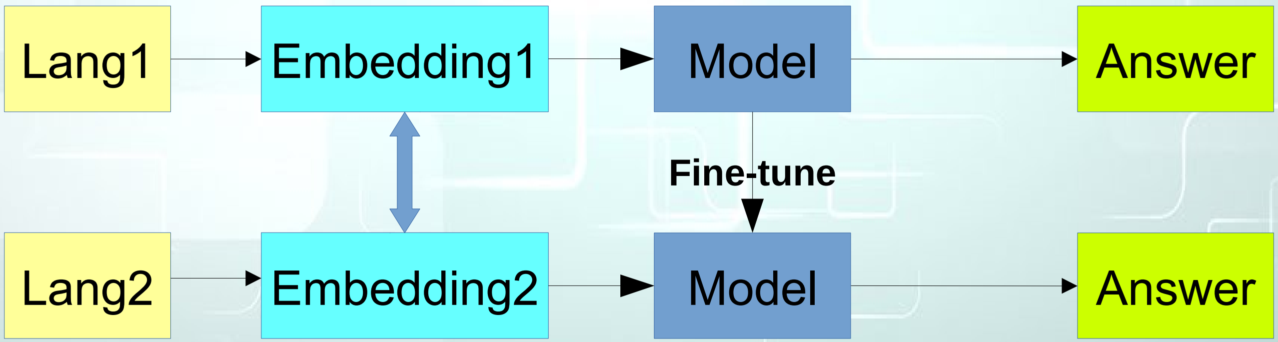
The IDEA

- Transfer Learning
  - Transfer the learned knowledge from a resource rich language (like English) to a resource poor language (like Spanish)

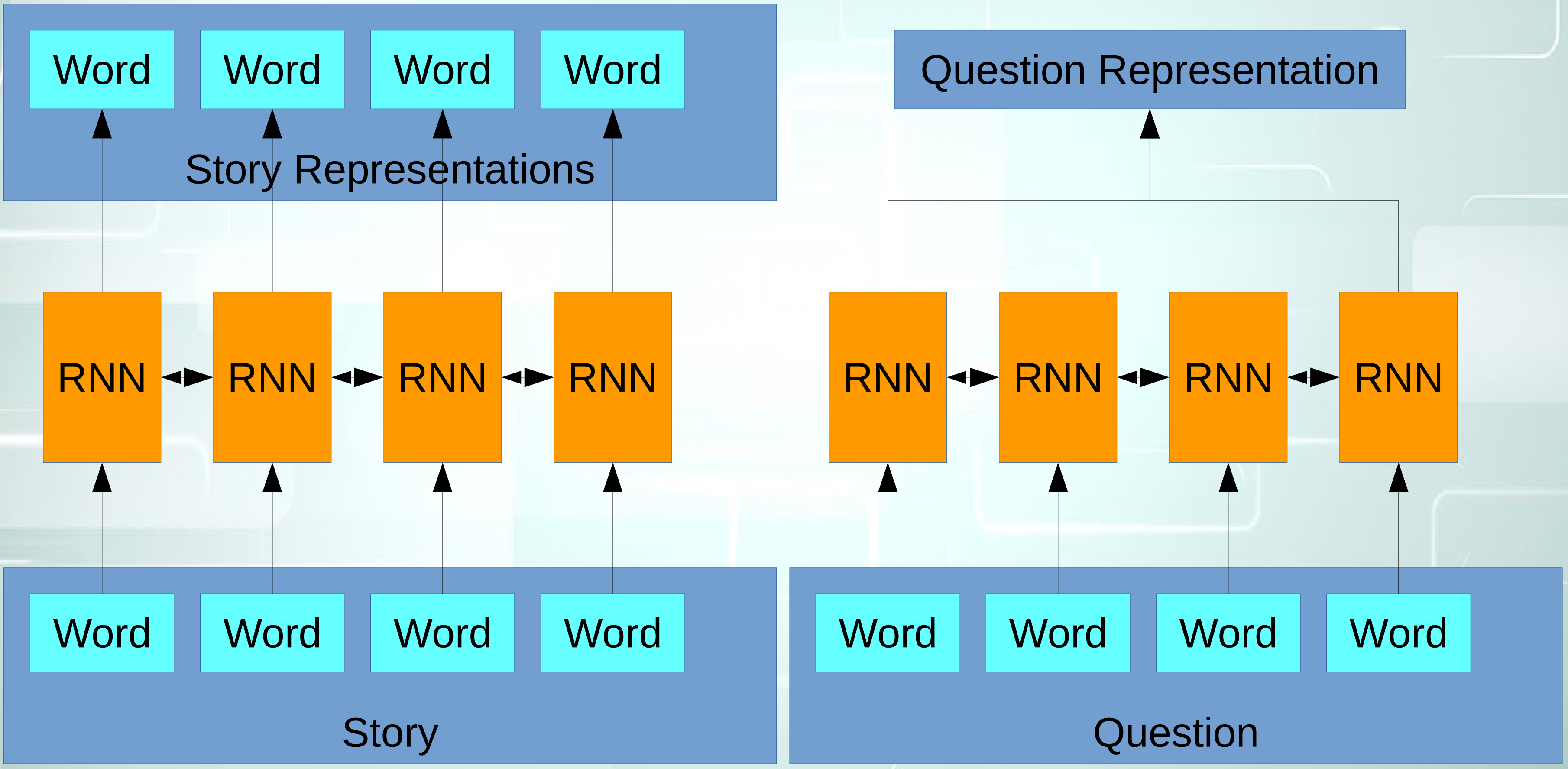
- Jointly learn the QA models for Engilsh & Spanish
  - Learn abstract knowledge which can be transferred across languages

Our Architecture

- Approach 1 – Direct Transfer
  - Train a model in Language 1, then fine-tune the model on a smaller dataset of Language 2
  - Assuming that some aspect of the model can be shared
  - To boost performance, we use aligned word embeddings

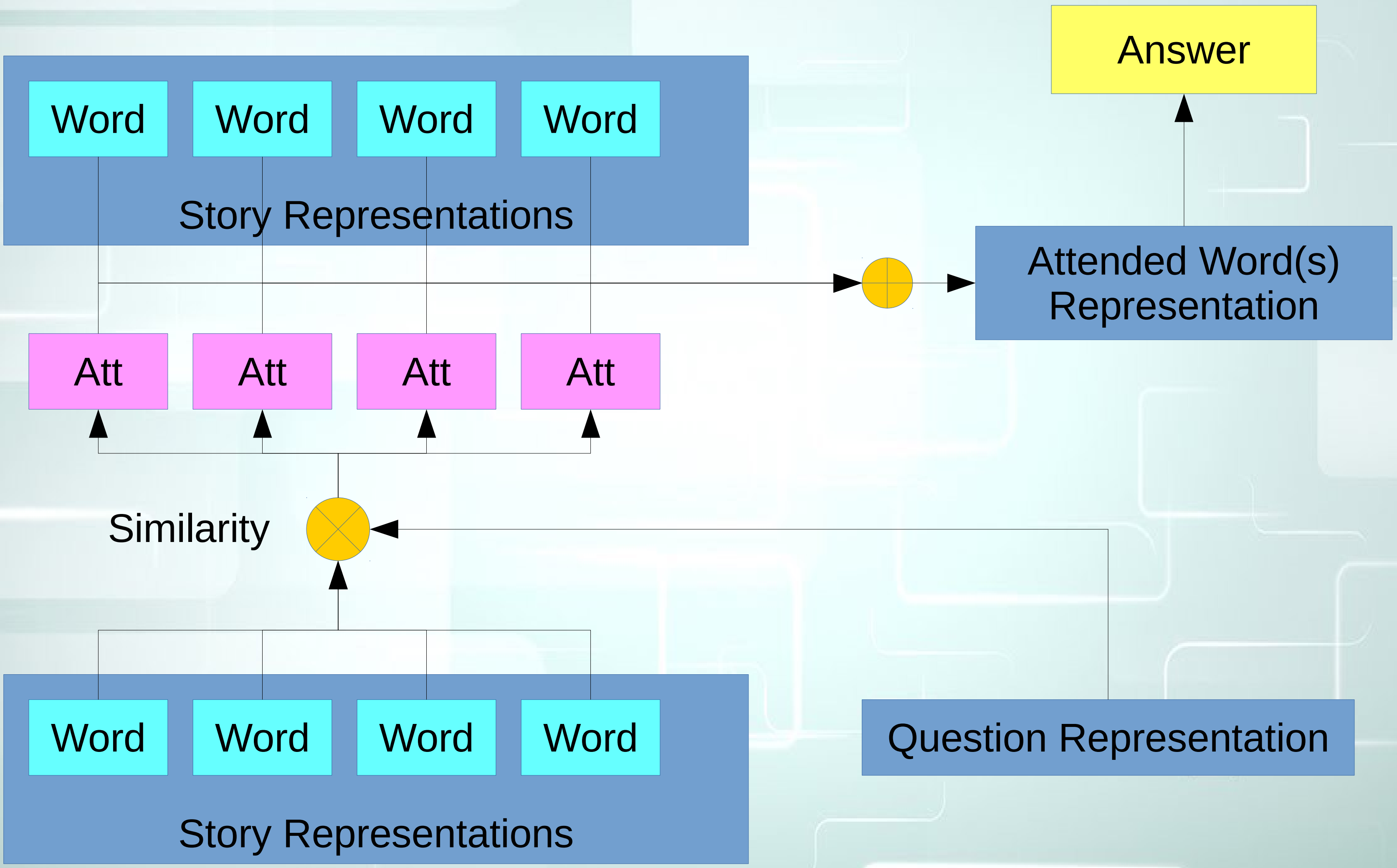


Our Architecture



Based on Hermann et al. (2015) and Chen et al. (2016)

Our Architecture



Based on Hermann et al. (2015) and Chen et al. (2016)

(@entity3) the @entity2 military claims to have shot down a @entity5 drone , state media reported tuesday . " @entity2 air defense systems shot down a @entity5 @entity9 (unmanned aerial vehicle) north of @entity11 , " the state - run @entity6 said . the @entity5 military lost contact...

@placeholder 's military claims
it shot down a @entity5 drone
north of @entity11

Match Context

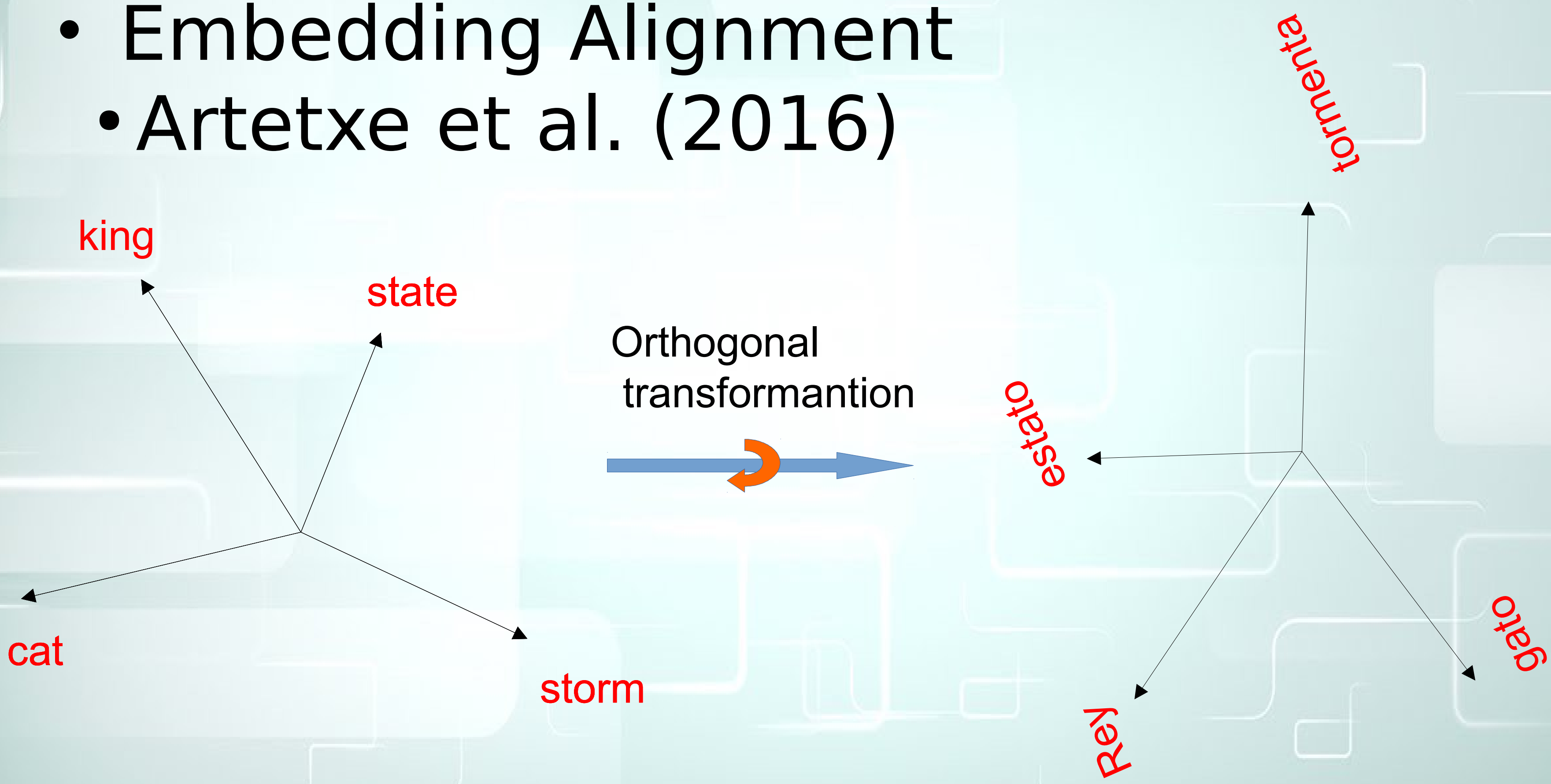
(@entity3) the @entity2 military claims to have shot down a @entity5 drone , state media reported tuesday . " @entity2 air defense systems shot down a @entity5 @entity9 (unmanned aerial vehicle) north of @entity11 , " the state - run @entity6 said . the @entity5 military lost contact...

Attend to key word

(@entity3) the **@entity2** military claims to have shot down a @entity5 drone , state media reported tuesday . " @entity2 air defense systems shot down a @entity5 @entity9 (unmanned aerial vehicle) north of @entity11 , " the state - run @entity6 said . the @entity5 military lost contact...

Our Architecture

- Embedding Alignment
  - Artetxe et al. (2016)



Our Architecture

- Embedding Alignment

Embedding Vectors of L1

a 0.32, 0.41, -0.55, …
the 0.02, -0.92, 0.33, ...

Embedding Vectors of L1

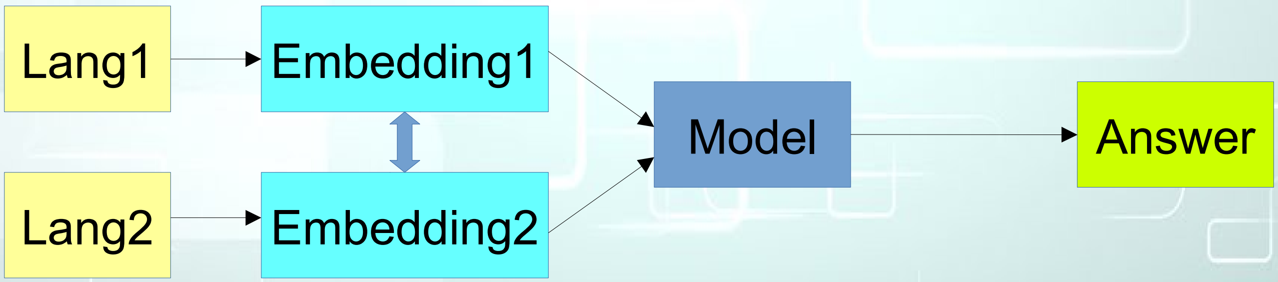
la 0.76, 0.39, -0.63, …
el 0.12, -0.92, 0.43, ...

Keyword Alignment:

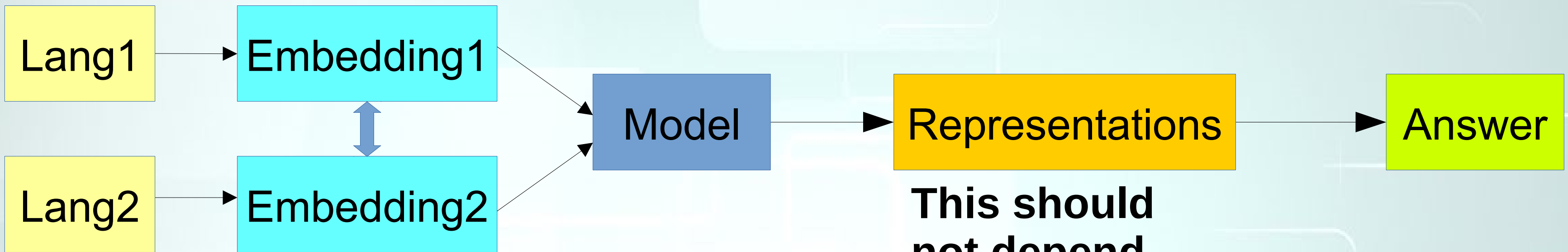
ions iones
iridium iridio
iris iris
irisated irisado
iron hierro
ironic irónico

Our Architecture

- Approach 2 – Joint Training
  - Simultaneously train the model on both languages
  - Challenge: different optimal weights for each language. Model might learn a bad “average” model
  - Proposed solution: use adversarial training to force the model to learn language-independent features



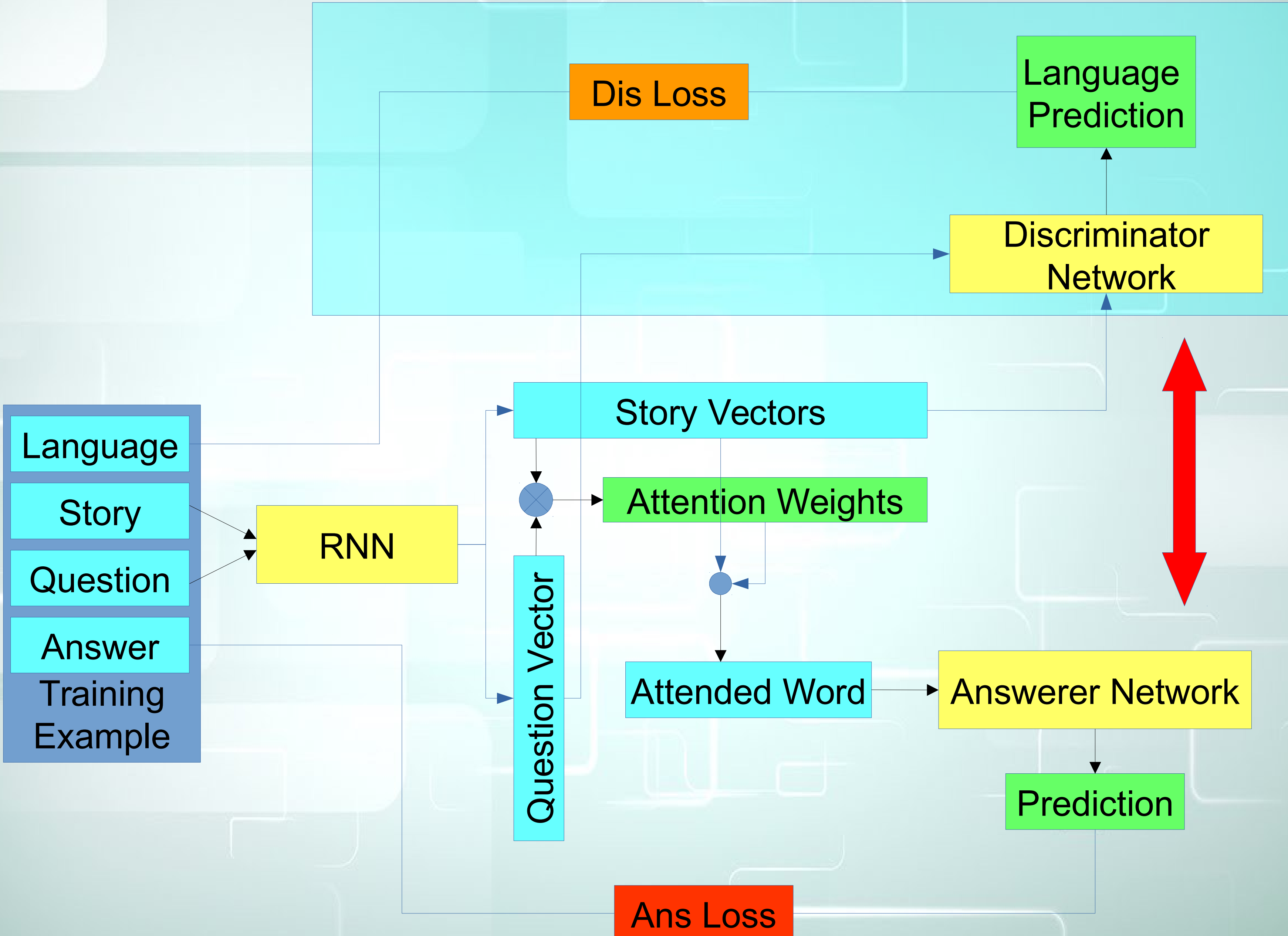
Our Architecture



This should not depend on the source language

Penalise language-dependency:

This should not be accurate



Dis Loss
Language
Prediction
Discriminator
Network
Story Vectors
Language
Story
Question
Answer
Training
Example
RNN
Attention Weights
Question Vector
Attended Word
Answerer Network
Prediction
Ans Loss

Training the Architecture

- Embedding Alignment Objective:

$$\underset{P}{argmin} \sum_{i} \left\| X_i P - Y_i \right\|$$ For all aligned words

- Adversarial Training Objective

$$\underset{\Theta_A, -\Theta_D}{argmin} \left[-logP(A|\Theta_A) + logP(L|\Theta_A, \Theta_D) \right]$$

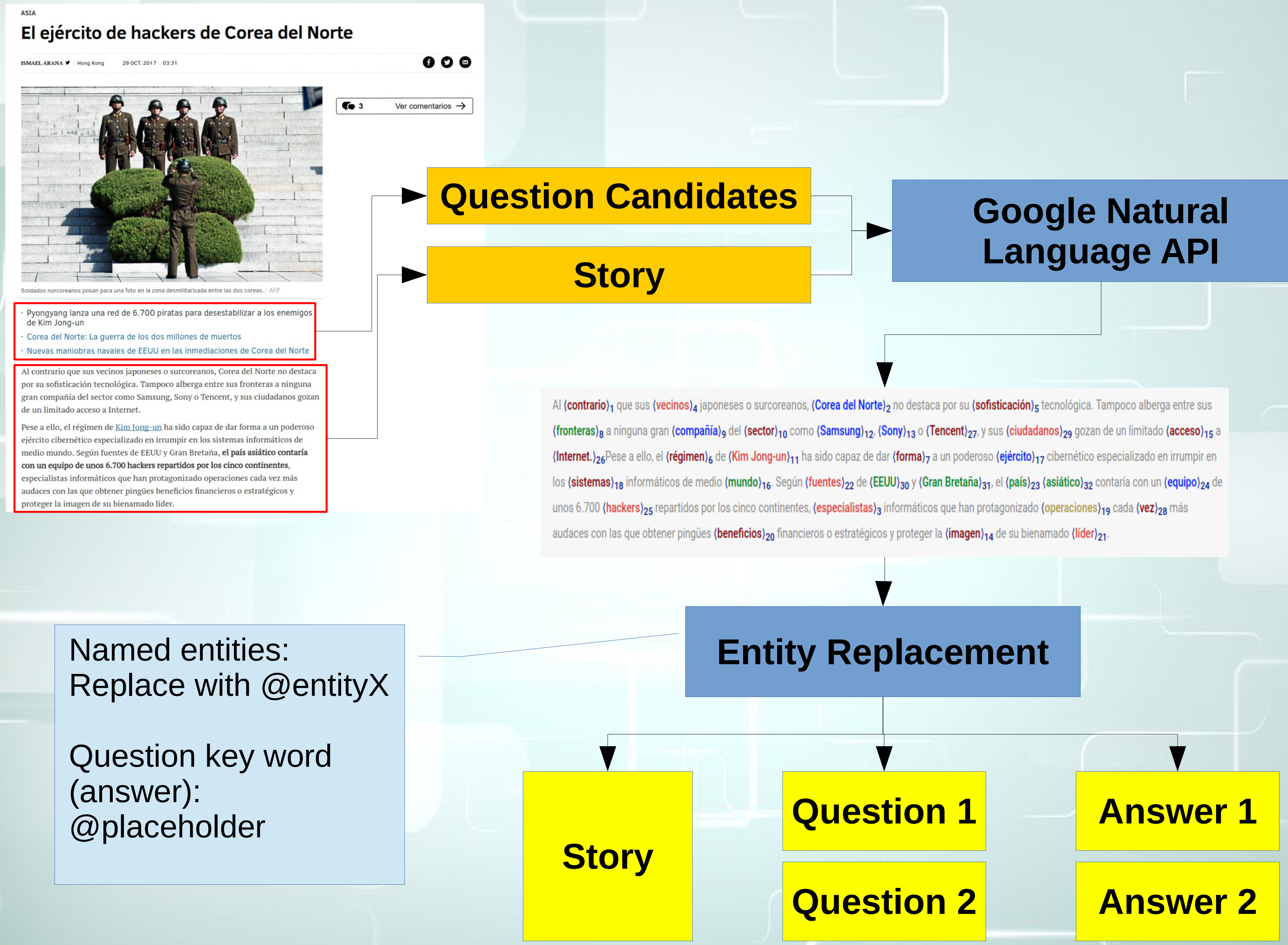
Follow the negative gradient of theta_D so that we minimise discrim. network loss wrt to it

Mimimise answerer network loss

Maximise discriminator Network loss

Experiments

- Datasets used:
  - English QA data
  - https://github.com/deepmind/rc-data/ (Hermann et al., NIPS 2015)
  - 380298 training QA pairs, 3924 dev and test QA pairs
  -
  - Spanish QA data
  - Collected from www.elmondo.es and CNN Spanish (via cached links on Wayback Machine)
  - Processed by Google Natural Language API (named entity recognition)
  - 69289 training QA pairs, 3839 dev and test QA pairs



ASIA
El ejército de hackers de Corea del Norte
ISMAEL ARANA | Hong Kong 29 OCT. 2017 | 03:31
3 Ver comentarios
Soldados norcoreanos posan para una foto en la zona desmilitarizada entre las dos coreas. / AFP
Pyongyang lanza una red de 6.700 piratas para desestabilizar a los enemigos de Kim Jong-un
Corea del Norte: La guerra de los dos millones de muertos
Nuevas maniobras navales de EEUU en las inmediaciones de Corea del Norte
Al contrario que sus vecinos japoneses o surcoreanos, Corea del Norte no destaca por su sofisticación tecnológica. Tampoco alberga entre sus fronteras a ninguna gran compañía del sector como Samsung, Sony o Tencent, y sus ciudadanos gozan de un limitado acceso a Internet.
Pese a ello, el régimen de Kim Jong-un ha sido capaz de dar forma a un poderoso ejército cibernético especializado en irrumpir en los sistemas informáticos de medio mundo. Según fuentes de EEUU y Gran Bretaña, **el país asiático contaría con un equipo de unos 6.700 hackers repartidos por los cinco continentes**, especialistas informáticos que han protagonizado operaciones cada vez más audaces con las que obtener pingües beneficios financieros o estratégicos y proteger la imagen de su bienamado líder.
Question Candidates
Story
Google Natural Language API
Al ⟨contrario⟩$_{1}$ que sus ⟨vecinos⟩$_{4}$ japoneses o surcoreanos, ⟨Corea del Norte⟩$_{2}$ no destaca por su ⟨sofisticación⟩$_{5}$ tecnológica. Tampoco alberga entre sus ⟨fronteras⟩$_{8}$ a ninguna gran ⟨compañía⟩$_{9}$ del ⟨sector⟩$_{10}$ como ⟨Samsung⟩$_{12}$, ⟨Sony⟩$_{13}$ o ⟨Tencent⟩$_{27}$, y sus ⟨ciudadanos⟩$_{29}$ gozan de un limitado ⟨acceso⟩$_{15}$ a ⟨Internet.⟩$_{26}$Pese a ello, el ⟨régimen⟩$_{6}$ de ⟨Kim Jong-un⟩$_{11}$ ha sido capaz de dar ⟨forma⟩$_{7}$ a un poderoso ⟨ejército⟩$_{17}$ cibernético especializado en irrumpir en los ⟨sistemas⟩$_{18}$ informáticos de medio ⟨mundo⟩$_{16}$. Según ⟨fuentes⟩$_{22}$ de ⟨EEUU⟩$_{30}$ y ⟨Gran Bretaña⟩$_{31}$, el ⟨país⟩$_{23}$ ⟨asiático⟩$_{32}$ contaría con un ⟨equipo⟩$_{24}$ de unos 6.700 ⟨hackers⟩$_{25}$ repartidos por los cinco continentes, ⟨especialistas⟩$_{3}$ informáticos que han protagonizado ⟨operaciones⟩$_{19}$ cada ⟨vez⟩$_{28}$ más audaces con las que obtener pingües ⟨beneficios⟩$_{20}$ financieros o estratégicos y proteger la ⟨imagen⟩$_{14}$ de su bienamado ⟨líder⟩$_{21}$.
Entity Replacement
Named entities: Replace with @entityX
Question key word (answer): @placeholder
Story
Question 1
Question 2
Answer 1
Answer 2

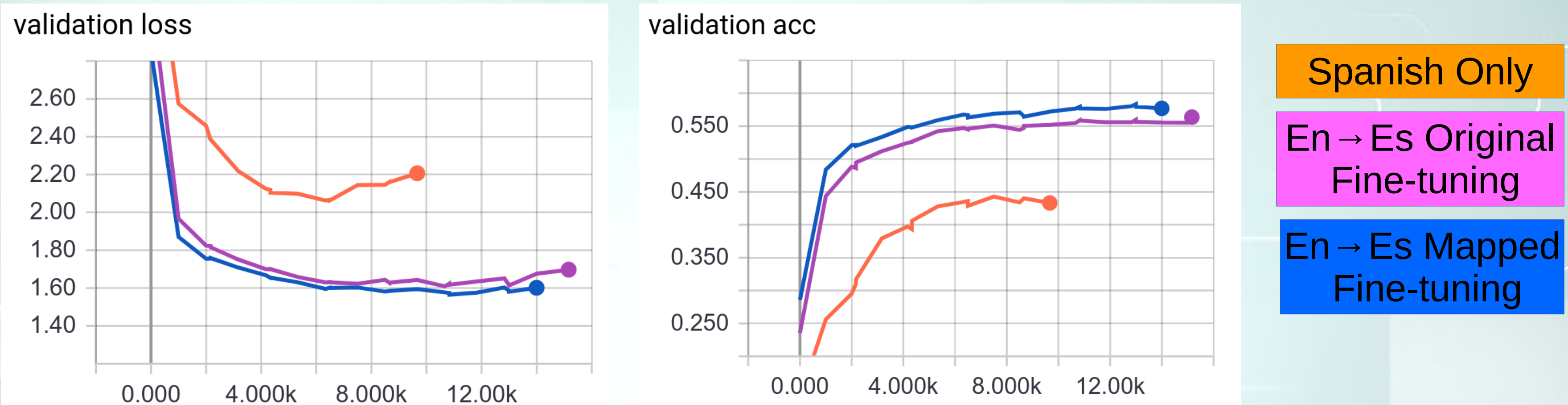
Experiments

- Embeddings used:
  - FastText by Facebook
  - https://github.com/facebookresearch/fastText/blob/master/pretrained-vectors.md
    (P. Bojanowski, E. Grave, A. Joulin, T. Mikolov, Enriching Word Vectors with Subword Information)
  - 300 dimension word vectors trained on Wikipedia text.
- Word Alignment Dictionary
  - http://opus.lingfil.uu.se/OpenSubtitles2012.php
    (Jörg Tiedemann, 2012, Parallel Data, Tools and Interfaces in OPUS. In Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC 2012))
  - Open Subtitles dataset
- Word alignment Algorithm
  - Artetxe et al. 2016

Results

	En.	Es.	En. → Es. (no training)	En. → Es. (training)
Individual Embeddings	0.65666	0.42089	0.21908	0.54538
Mapped Embeddings	0.66291	0.41596	0.26405	**0.56622**
Joint Training	0.63226	0.51001	-	-
Adversarial Training	**0.66979**	0.5133	-	(0.51714)

Analysis



1. The spanish dataset alone is not large enough to learn the context matching and entity mapping task

2. Initialising training on the Spanish dataset with the model trained on the English dataset helps accelerate training and improve accuracy on the Spanish dataset

3. Using aligned embeddings instead of individual embeddings slightly improves Spanish dataset performance

Analysis

- Using adversarial training improves performance for both languages individually
- The improvement on Spanish dataset is not as significant as using mapped embeddings + fine-tuning
  - Discussion next
- Surprisingly, the performance on English dataset is better than without adversarial training
  - It is likely that joint training and adversarial training provides some extra information while also acting as a form of regularisation

Analysis

- Why is adversarial training not as effective for Spanish dataset?
  - Model is “overwhelmed” by English data
    - But “supersampling” Spanish data does not improve performance while introducing overfitting
  - This particular task involves mainly context matching, which is already language-independent to some degree
    - Plus “token variable” (@entityX) representations are already shared
    - Attention mechanism is trained to focus on these token words
    - Adversarial training does not contribute much to the language-independence of answer representation
  - Truly language-independent features are difficult to learn
    - Word sense is sensitive to how it relates to other words in a specific language

Conclusions

- Cross-lingual knowledge transfer between QA models is possible
- Fine-tuning is effective when the task itself is not highly language-dependent
- Aligned embeddings generated with a limited dictionary can potentially improve cross-lingual tranfer learning performances
- Joint learning and language-independent represenation learning through adversarial training is promising, but not necessarily better than fine-tuning

Future Work

- Evaluate cross-lingual tranfer learning performance on QA tasks without entity replacement (i.e. more language-dependent)

- Use adversarial learning to fine-tune word embeddings

- Explore different options to represent “variables” (such as @entityX or @placeholder in this problem) within a text

- Evaluate transfer learning on bilingual (mixed) texts