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

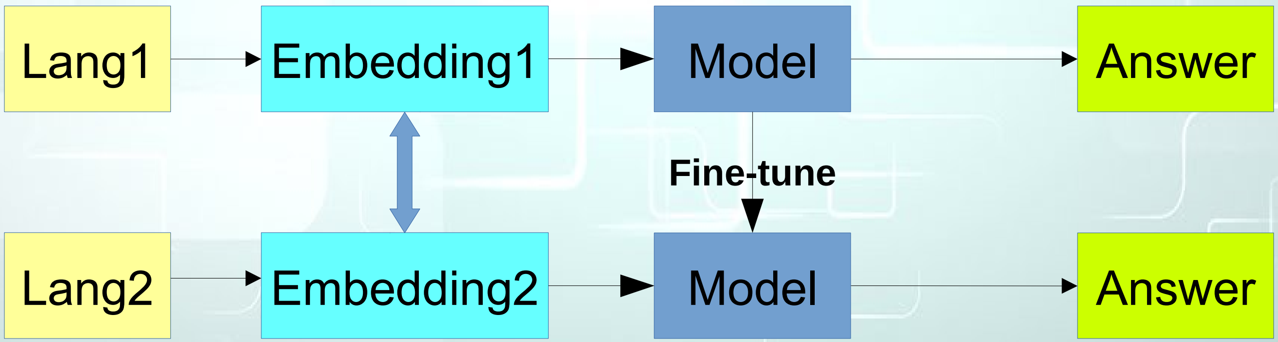
- Large amount of data is needed to train a neural QA system
  - 380K Q&A pairs in the CNN news dataset (machine-generated)
  - 120K Q&A pairs in the Maluuba NewsQA dataset (human-annotated)
  - 100,000+ Q&A pairs in SQUAD dataset
  - Such large datasets (especially human-annotated ones) are difficult 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

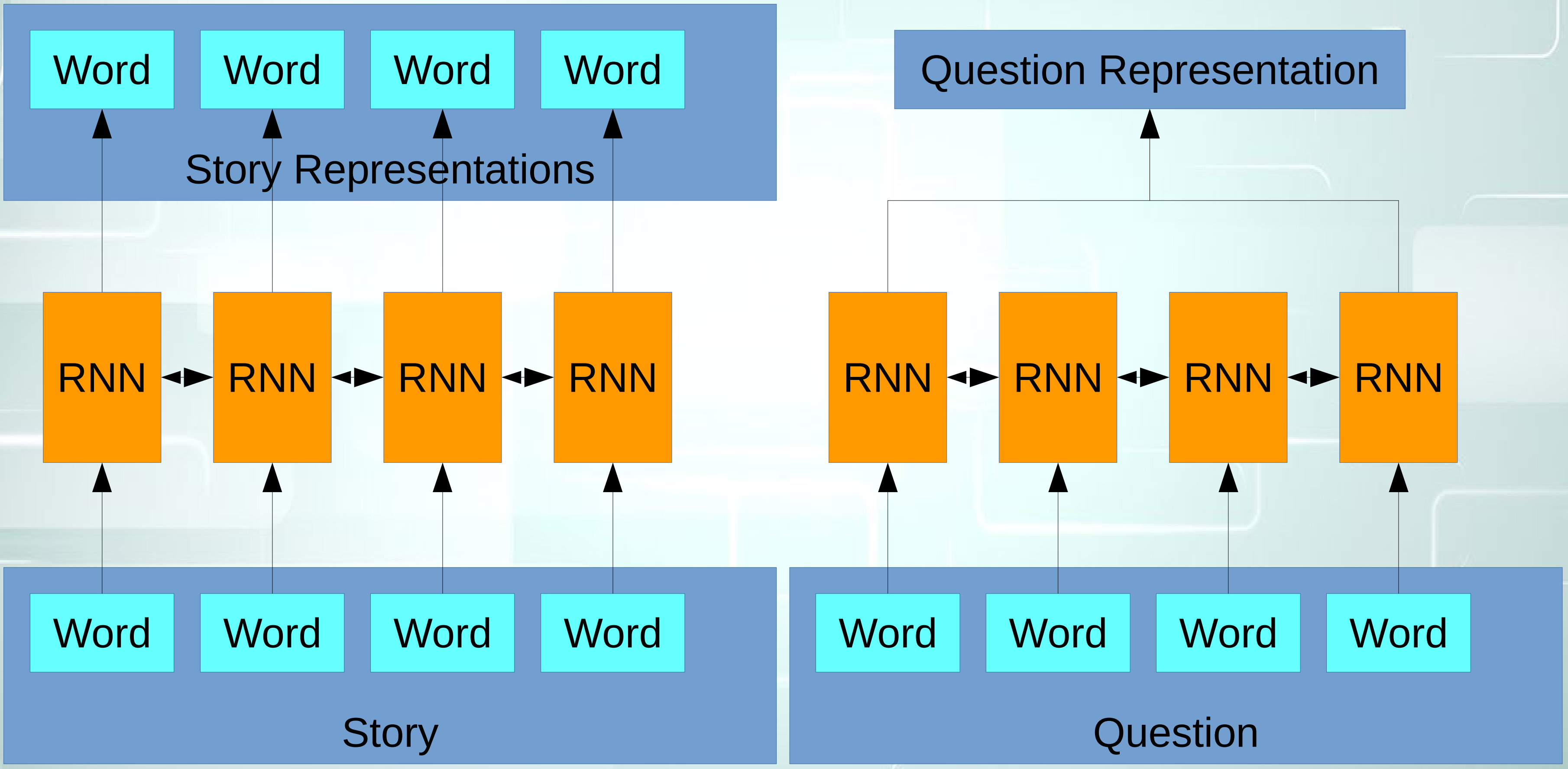
- Transfer Learning
  - Transfer the learned knowledge from a resource rich language (like English) to a resource poor language (like Spanish)
- Jointly learn the QA models for Engilsh & Spanish
  - Learn abstract knowledge which can be transferred across languages

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

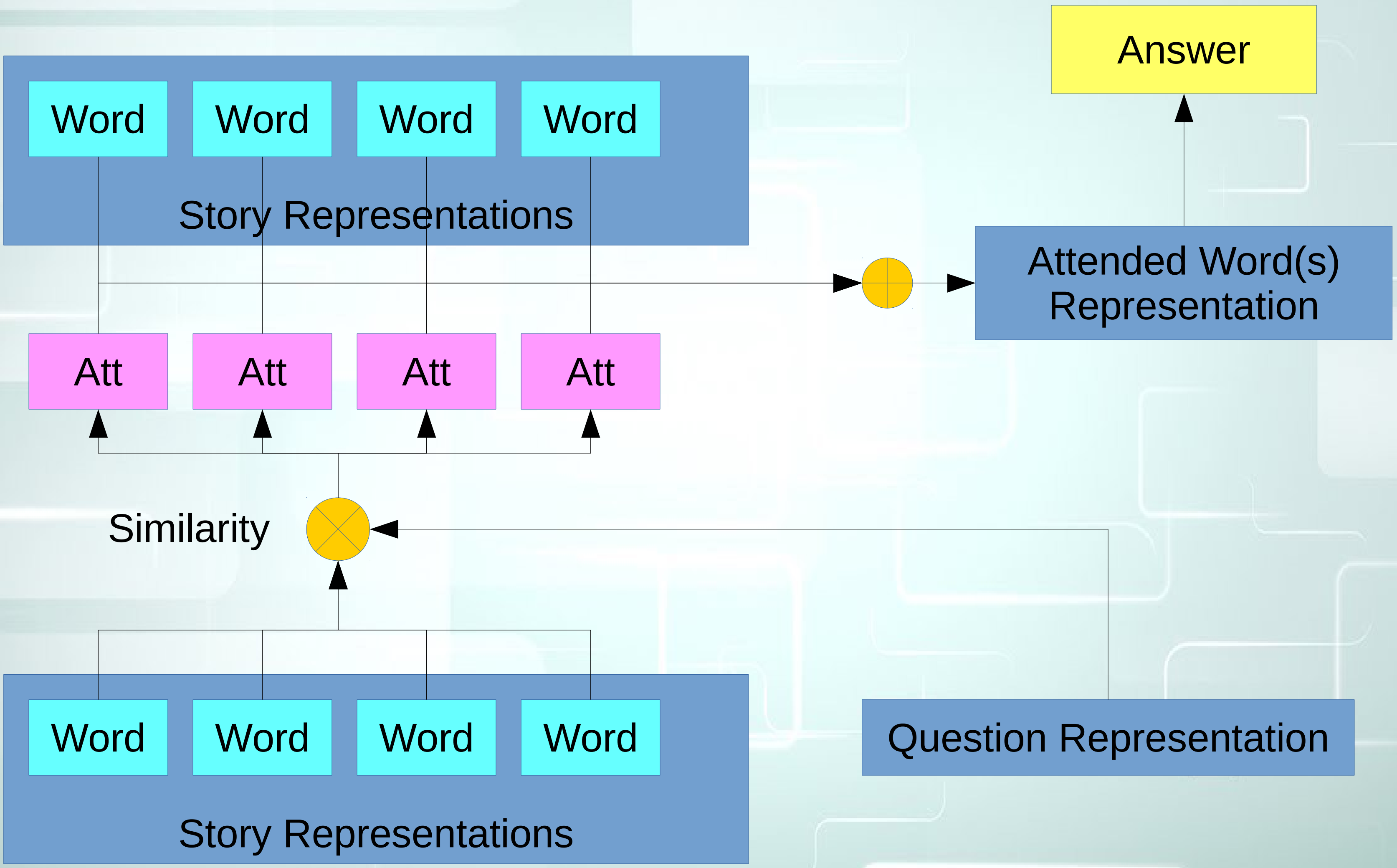


Our Architecture



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

Our Architecture

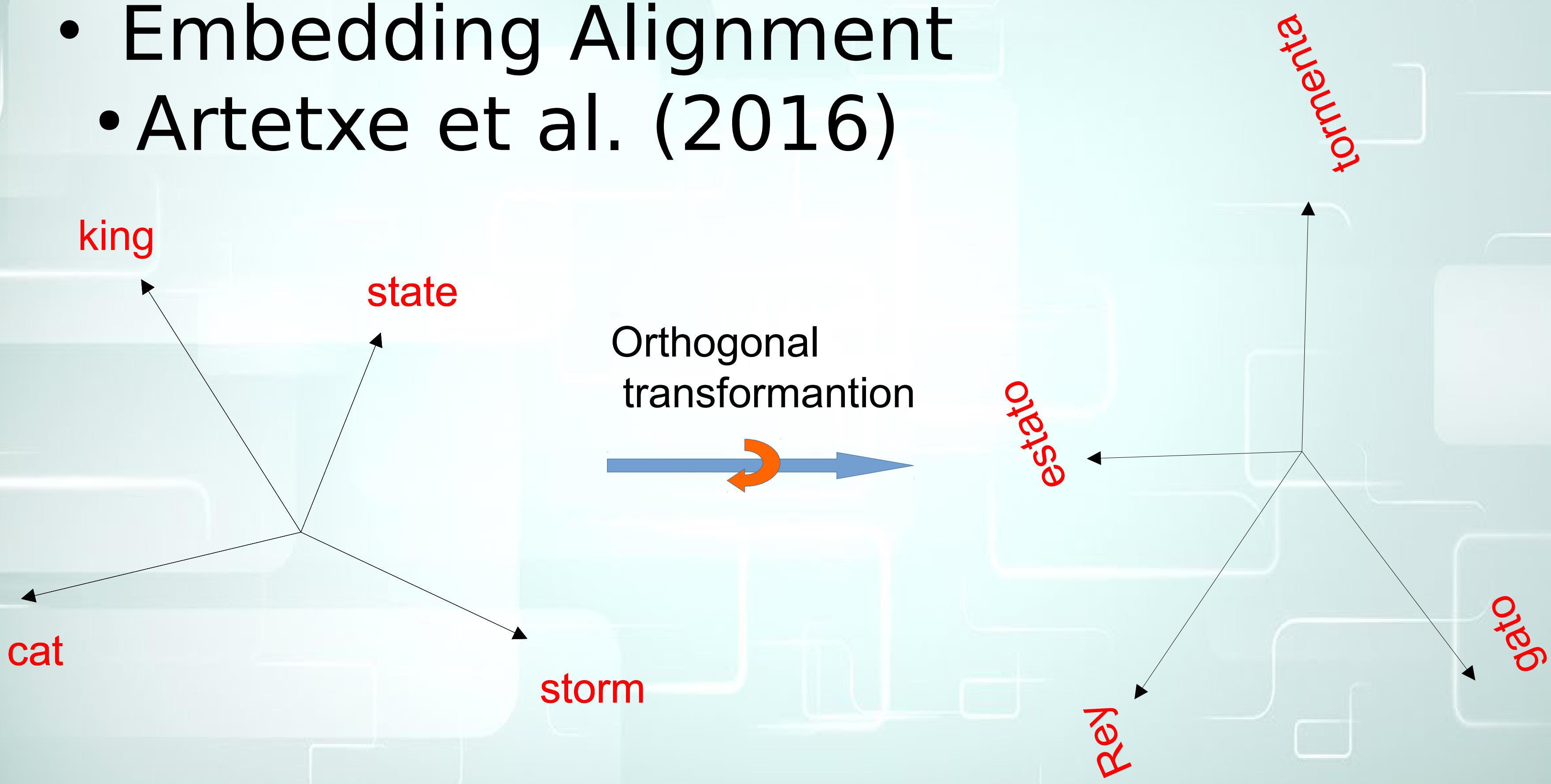


Based on Hermann et al. (2015) and Chen et al. (2016)


(@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

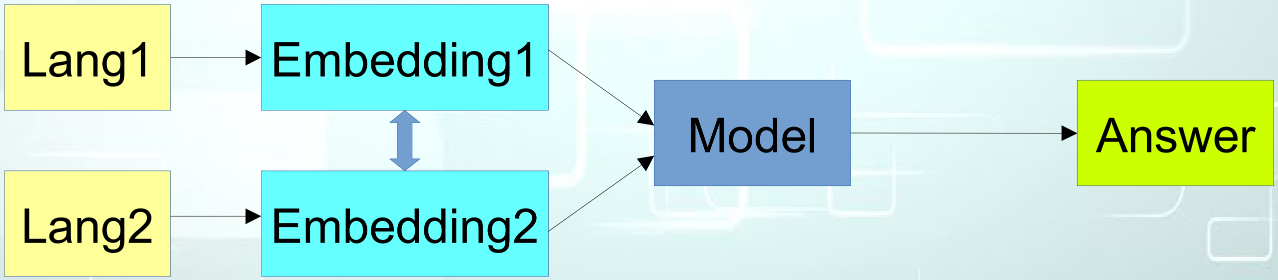
la 0.76, 0.39, -0.63, …
el 0.12, -0.92, 0.43, ...

Keyword Alignment:

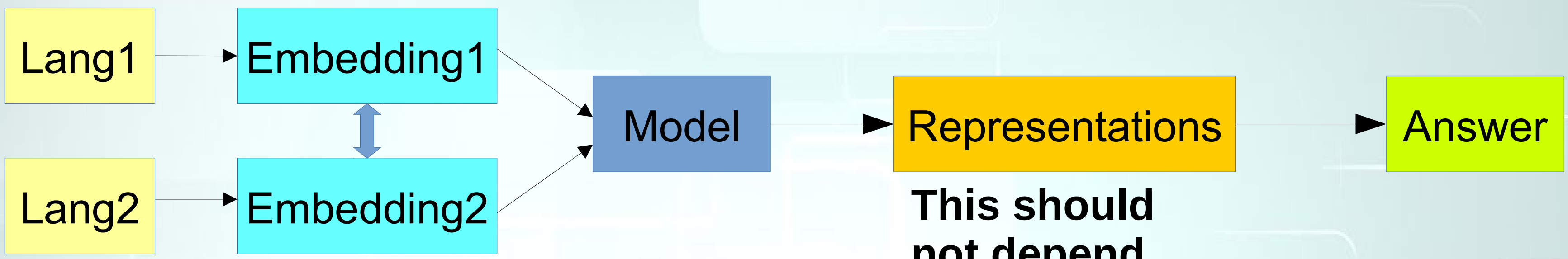
ions iones
iridium iridio
iris iris
irisated irisado
iron hierro
ironic irónico

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



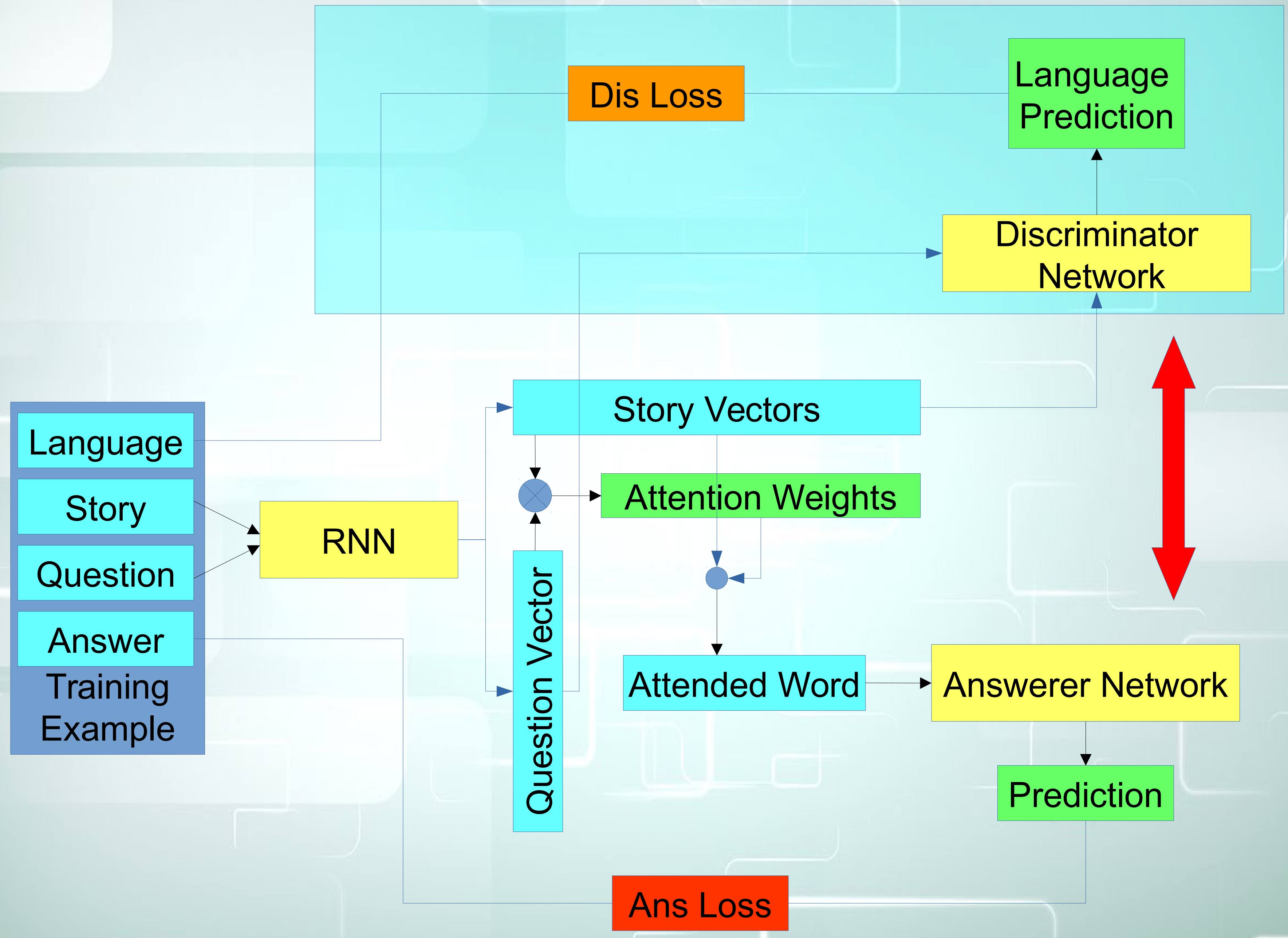
Our Architecture



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

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

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

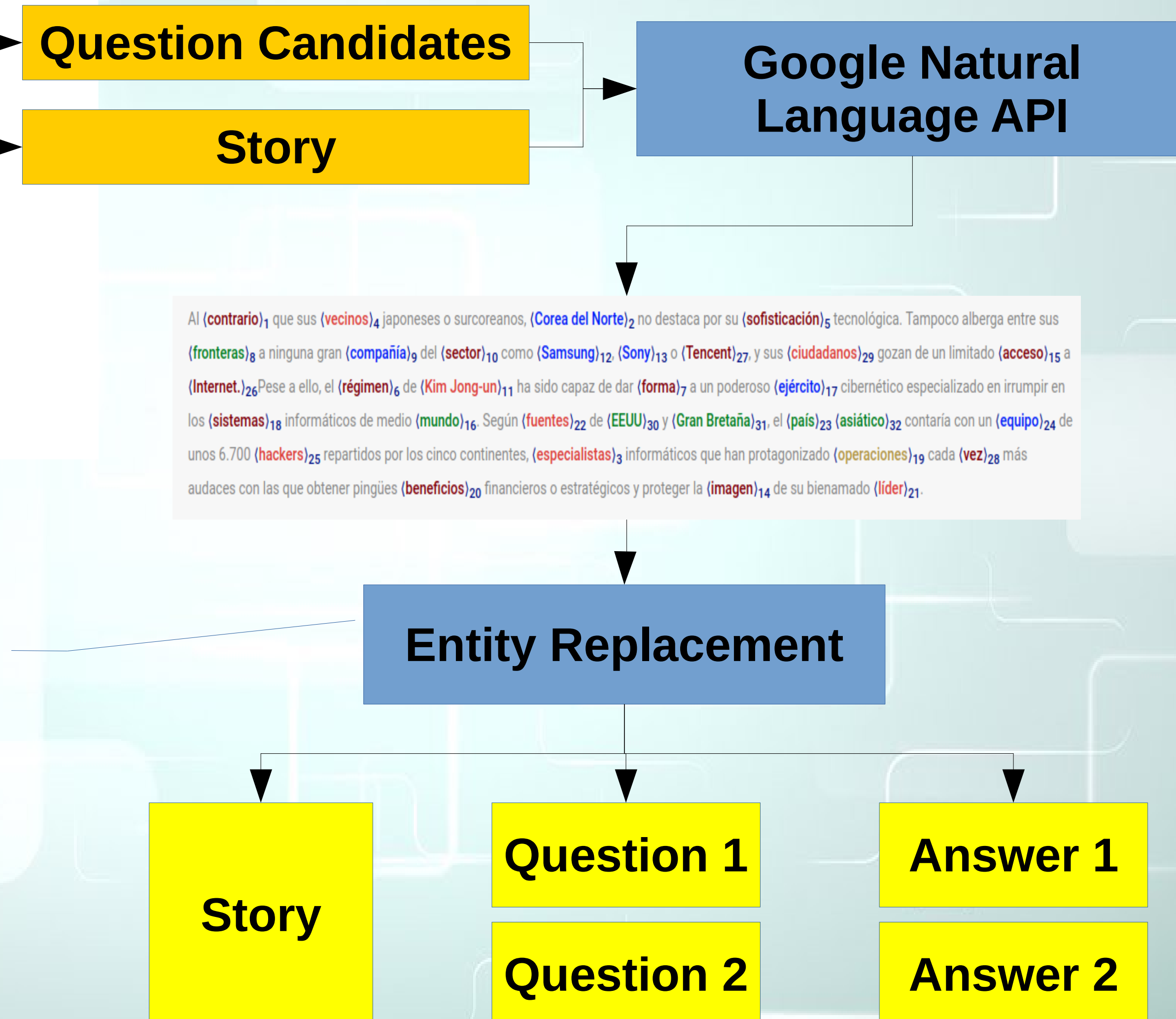
3 Ver comentarios

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.



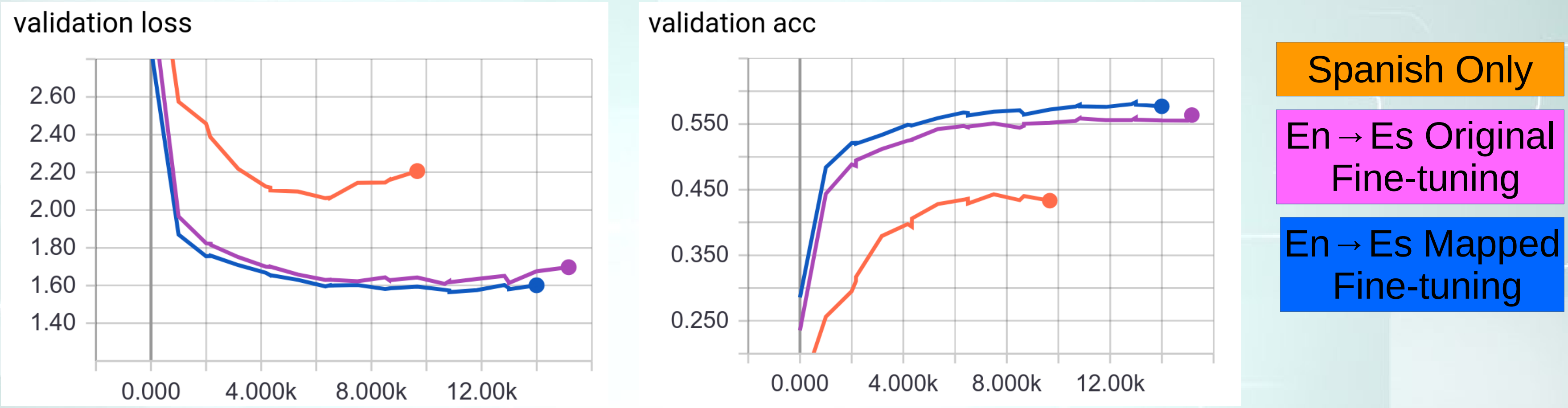
Experiments

- 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

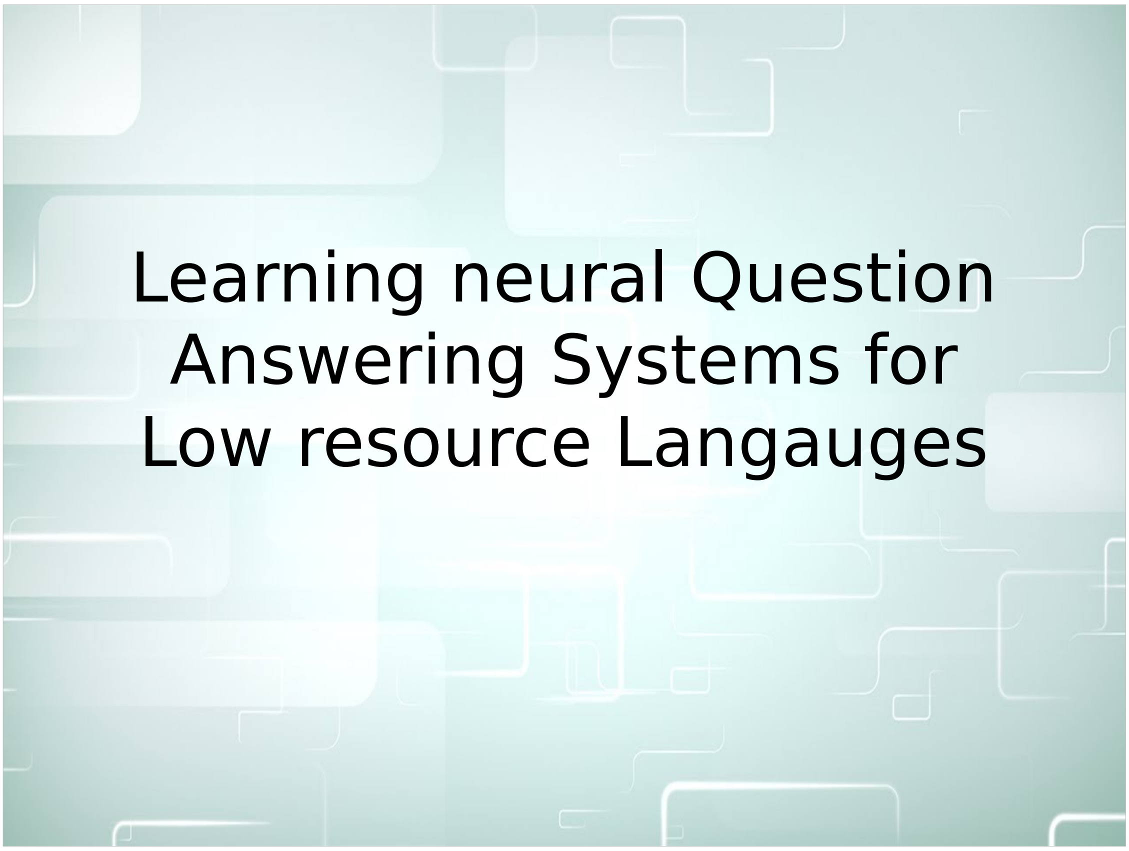
- Why is adversarial training not as effective for Spanish dataset?
  - Model is “overwhelmed” by English data
    - But “supersampling” Spanish data does not improve performance while introducing overfitting
  - This particular task involves mainly context matching, which is already language-independent 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



Welcome

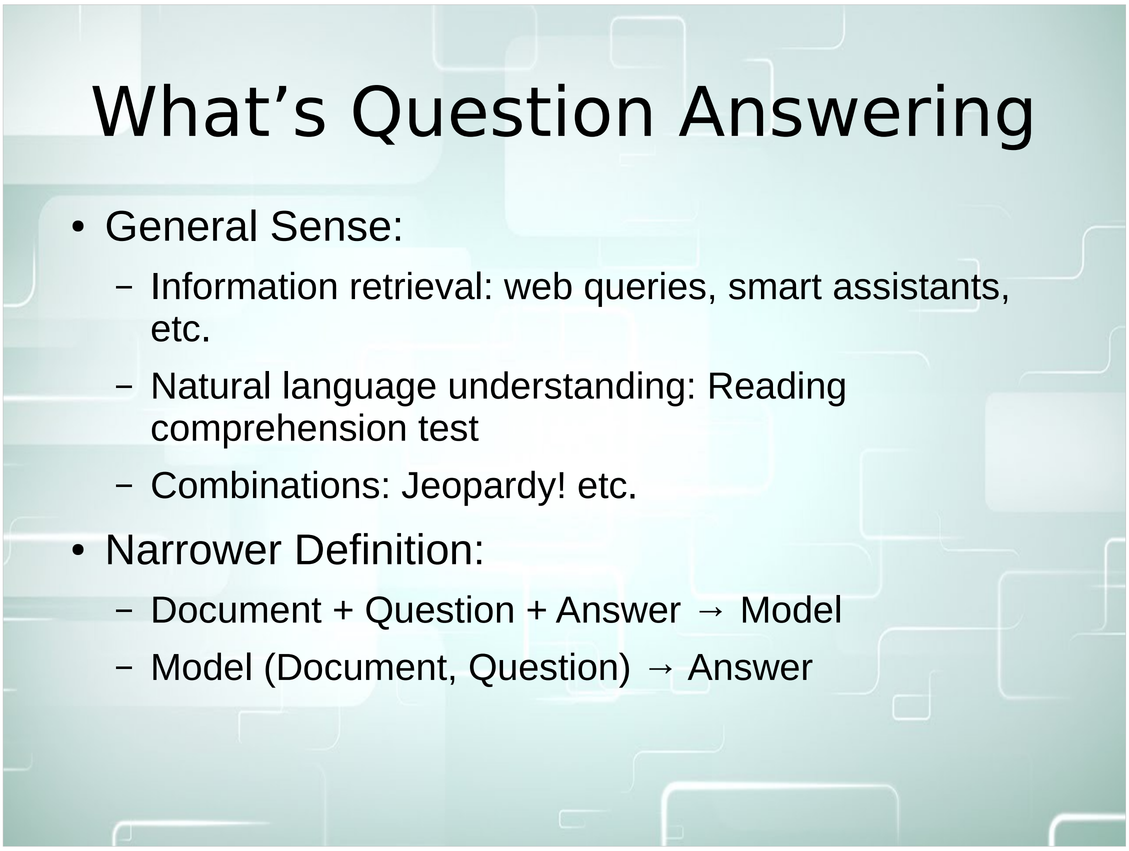
Topic: learning a neural question answering system that can transfer knowledge learned from resource-rich languages to relatively resource-poor languages

Breakdown:

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

Overview

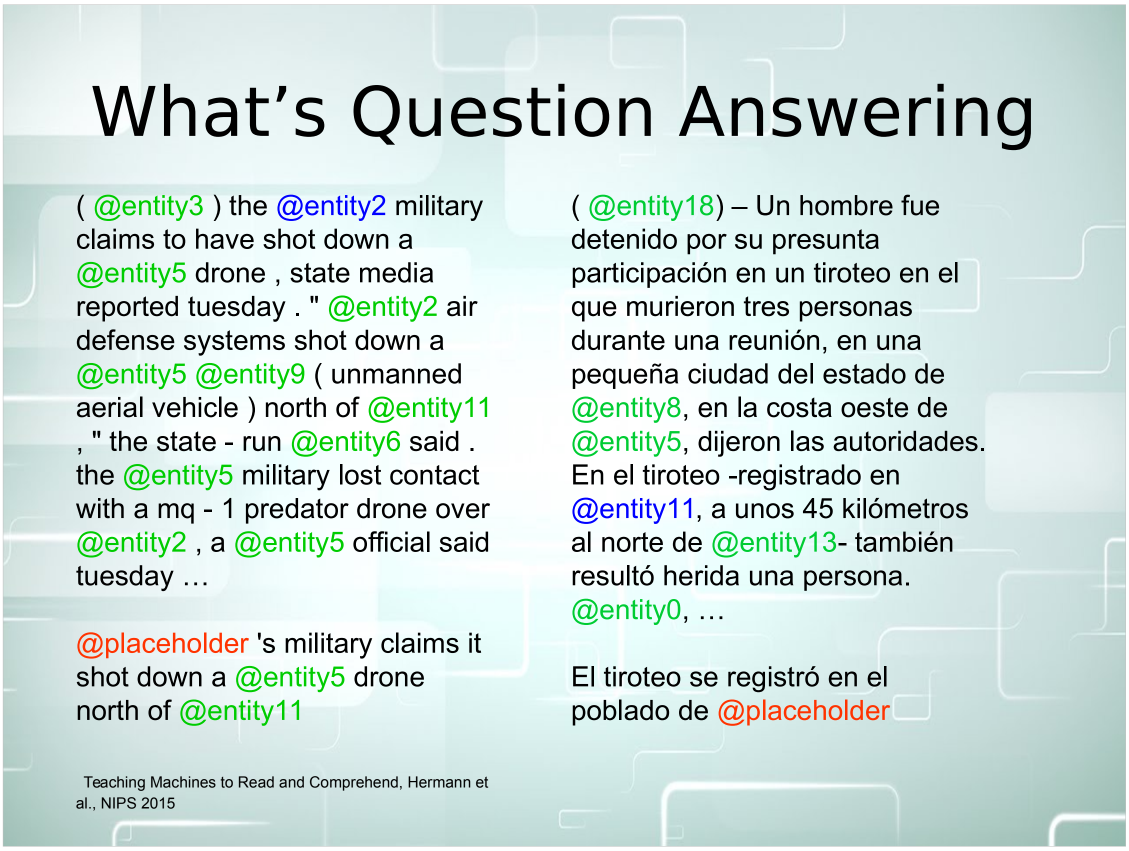
1. Problem statement: what is question answering
2. Neural question answering and the challenges for resource-poor languages
3. Proposed approaches to transfer learning
4. Experiment design
5. Experiment Results
6. Conclusions



“Question answering” can be understood in many different senses

…

But here we are interested in a specific form of QA (simple reading comprehension)

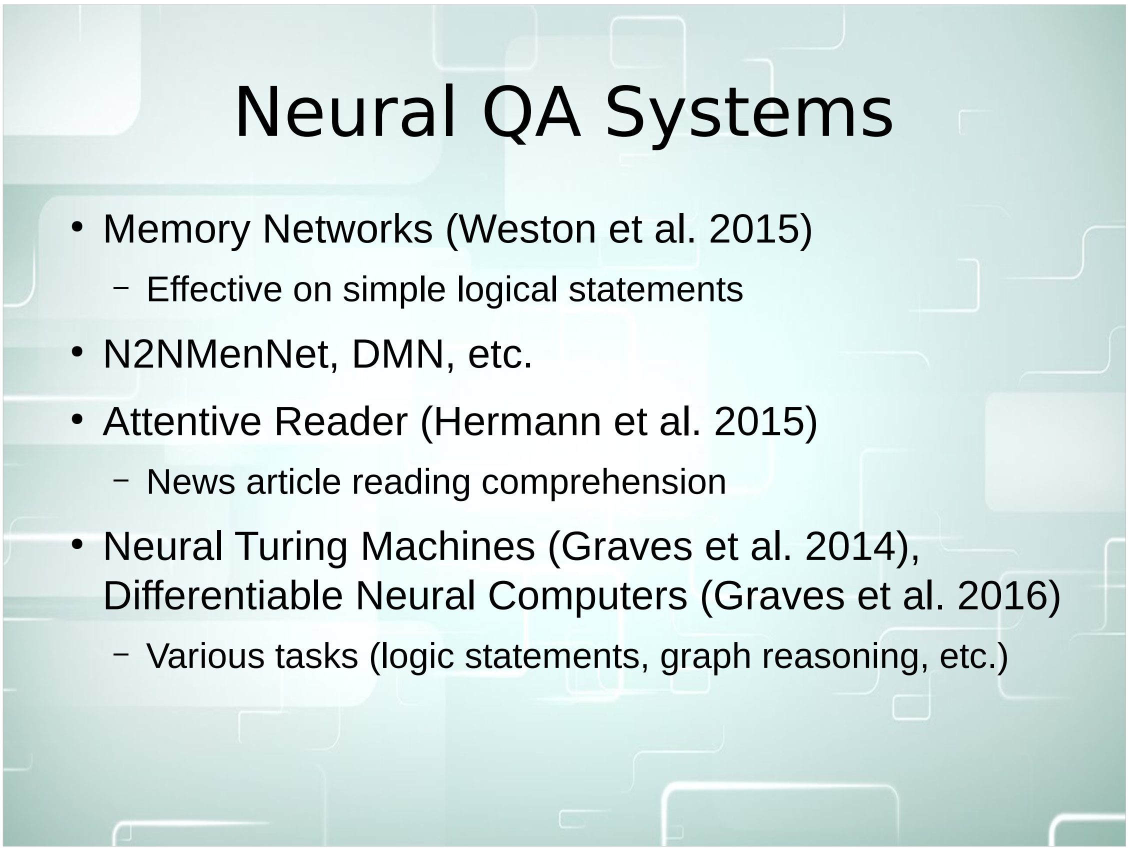


One example of a well-formulated QA problem is like this one purposed in Hermann et al. 2015 paper. Also the problem we will be focusing on later.

Story + answer format

Look for the best word in the story corresponding to the placeholder token

Named entities are replaced in this example. Reason – force the system to learn basic reasoning instead of simple collocations



(BRIEF multiple systems with different architectures to solve different types of problems)

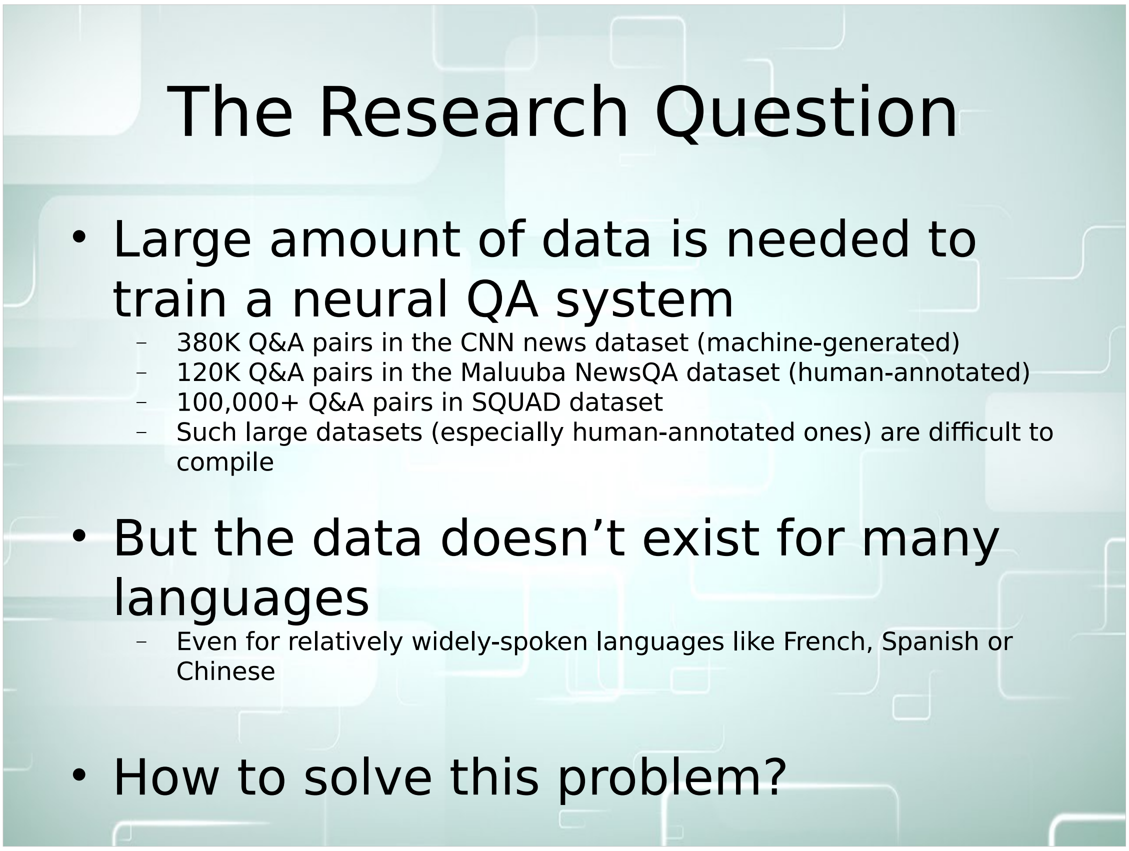
In recent years, many developments involving applying neural network-based systems to QA problems. Here are some notable examples.

Memory network model uses an external memory module and attention-based memory access to effectively solve simple logical deduction problems

Extensions such as N2NMemNet, Dynamic MN etc.

Attentive Reader achieves reasonable results on news article QA problems

NTM/DNC able to handle a variety of tasks



Thst is all good, however…

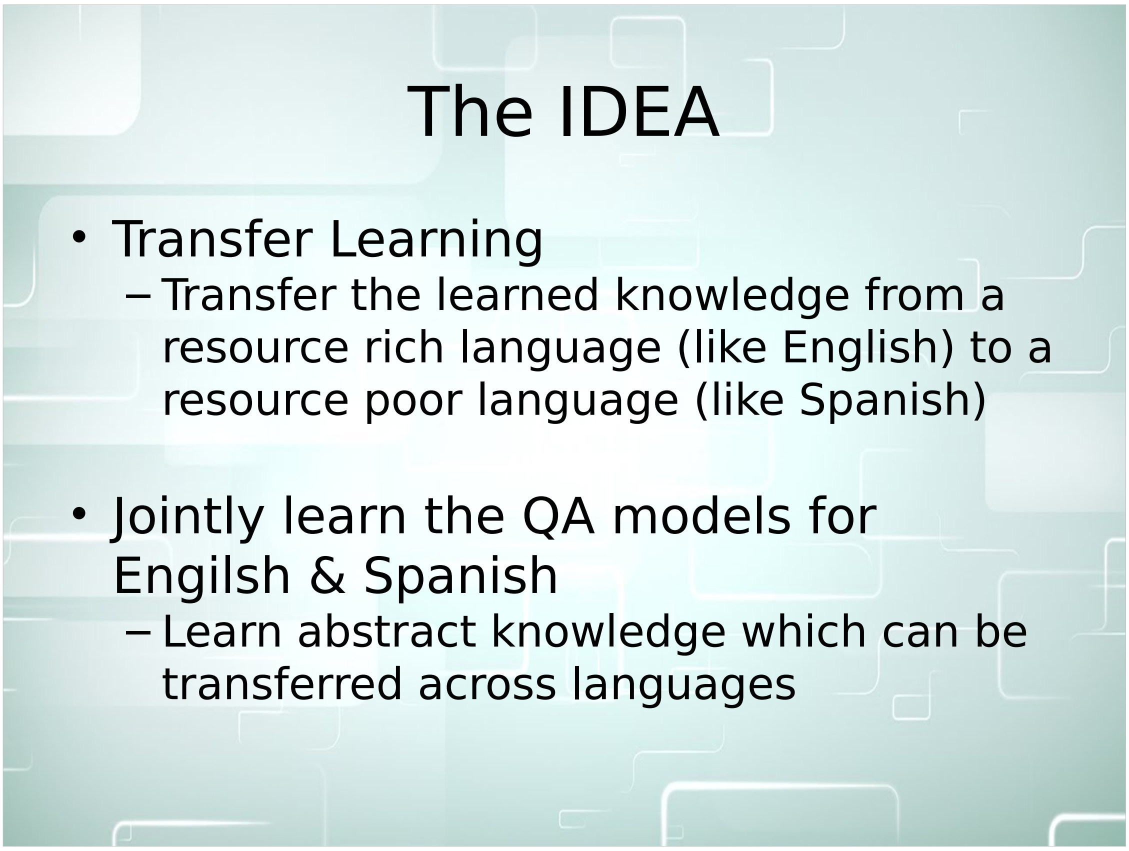
As with most neural networks, training a neural network-based QA system requires a large amount of training data, some of which are rather difficult to come by…

Here are some of the datasets used in previously mentioned research or current research

Especially difficult if data have to be human-annotated

Current status: relative abundance of datasets for English, but much less available data for other languages

Challenging for developing systems on new languages for which we cannot easily obtain high-quality high-volume datasets



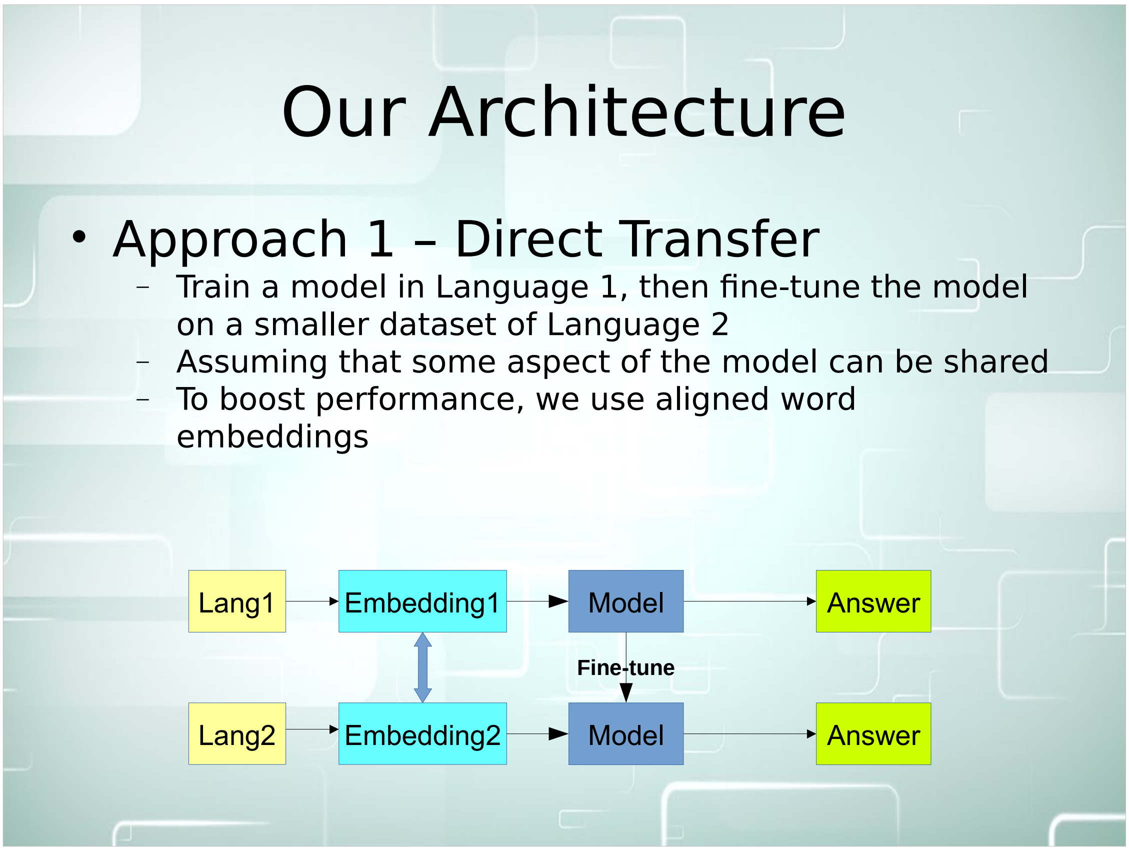
We wish to use the knowledge learned by models trained on English data to help models to be trained on other languages.

Even though the questions and answers are in different languages, the model may learn either some form of generic linguistic knowledge and exploit the similarity between two languages, or learn some task-specific but language-indepdent knowledge such as how to perform context matching.

Generally two basic approaches to tranfer learning: fine-tuning or joint training

This can be achieved in two ways:

1. sequential transfer learning
2. Joint learning (simultaneous)

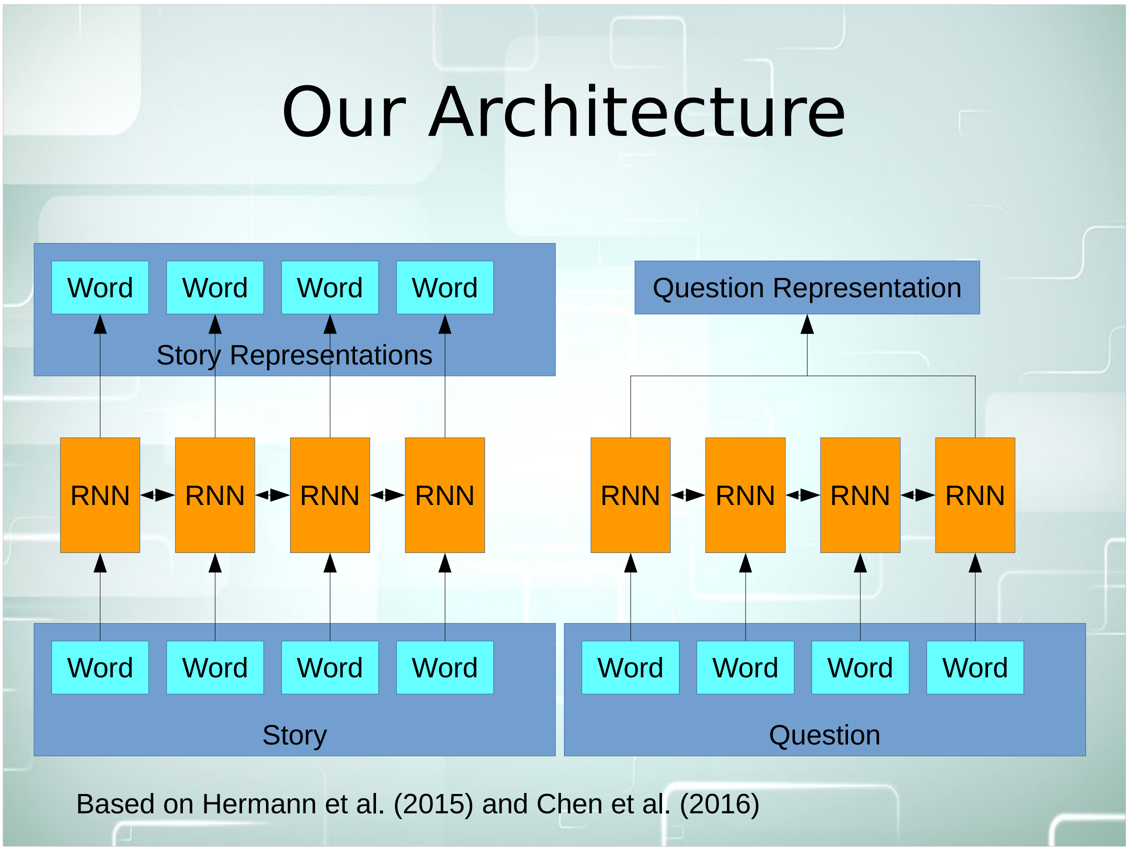


Here we propose two network architectures for the transfer learning problem. One based on direct tranfer learning and another only joint learning.

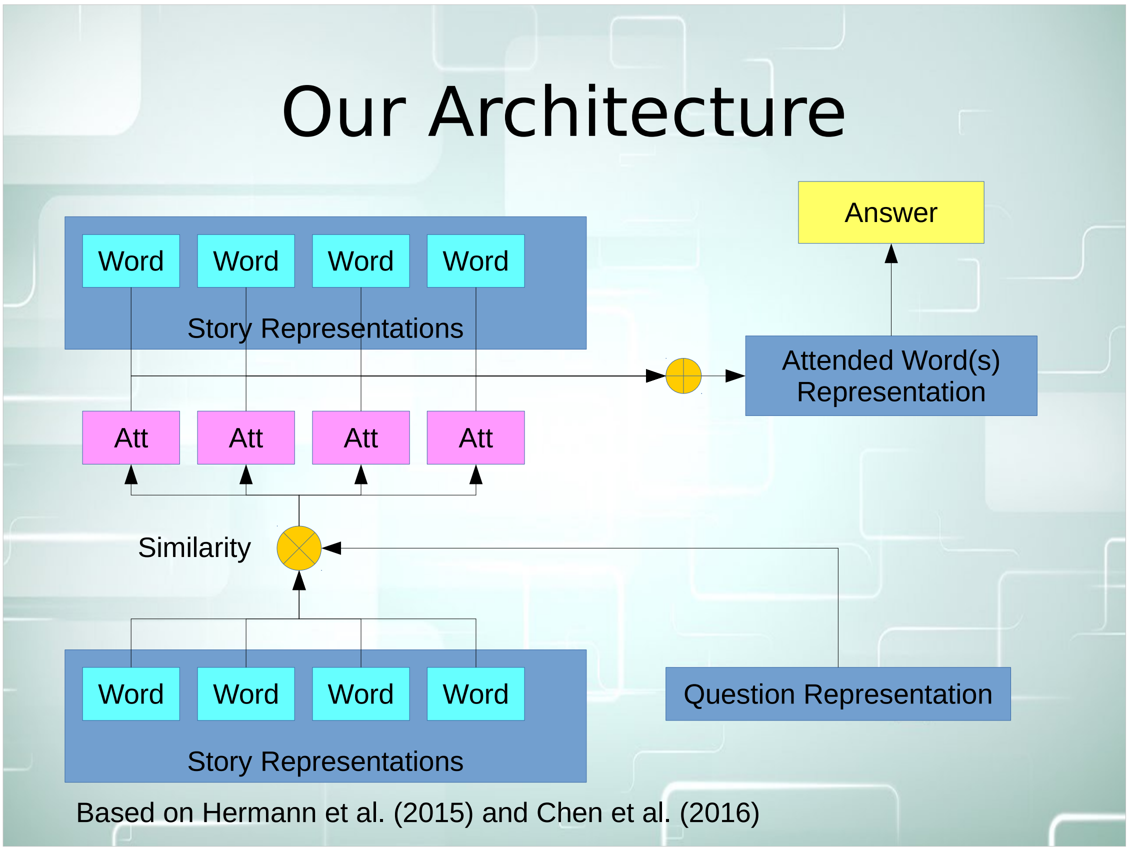
In this first approach we attempt to build a model in Lang 1 first, then fine-tune the parameters using training examples of Lang 2

We use fine-tuning to allow the model to “inherit” knowledge learned from Lang 1 and adapt it to Lang 2

Aligned word embeddings – discuss later



Start with story and question words
Convert to embeddings
Run through bidirectional RNNs respectively to obtain a context-depedent representation of words in the story and a summary representation of the question



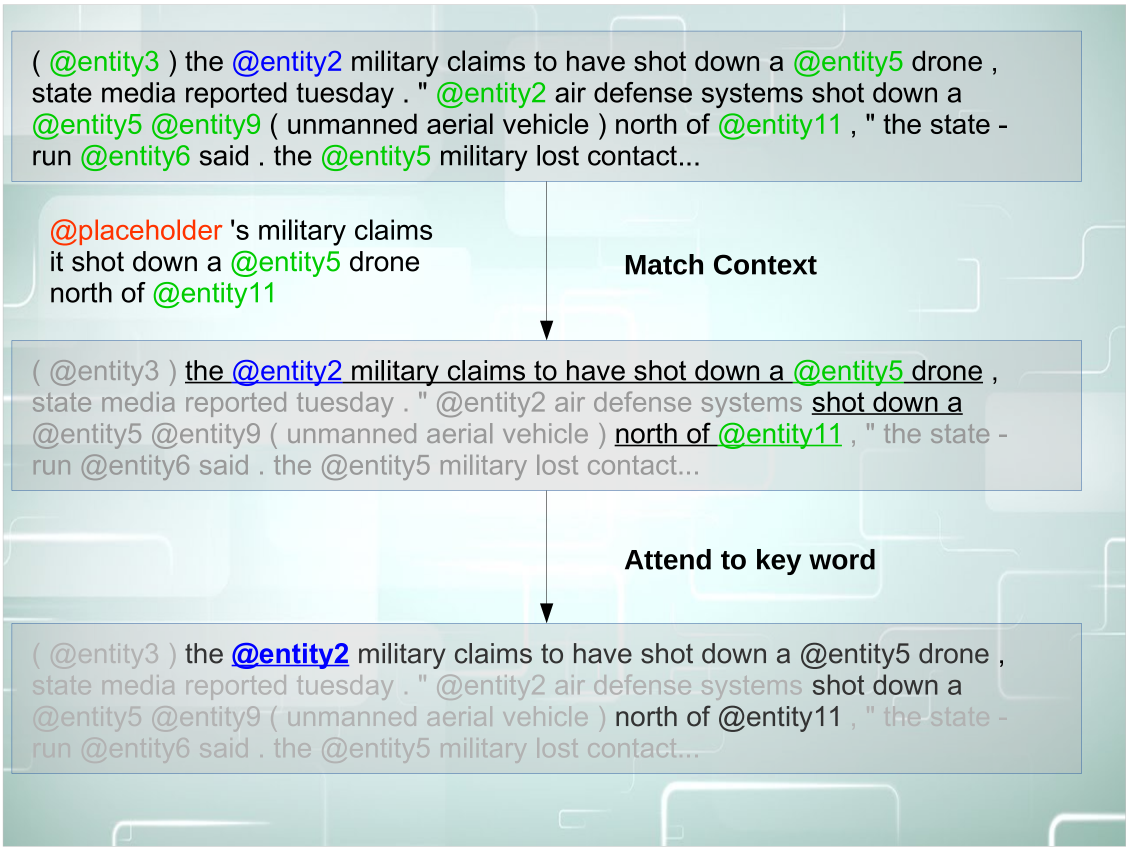
Use attention mechanism to focus on part of story:

We calculate a similarity measure between the question representation and the (contextual) representation of each word in the story and define that as the attention for each word

Sum of contextual story word representations weighted by attentions as the attended word(s) representation

Encodes what word or words are focused on by the network

Answerer network to map the attended word represenation to an answer token id



Example of how the network works

First step – calculate representations of words and question

Second – use similarity measure to compare the question vs story and match possible context(s) that contains the answer

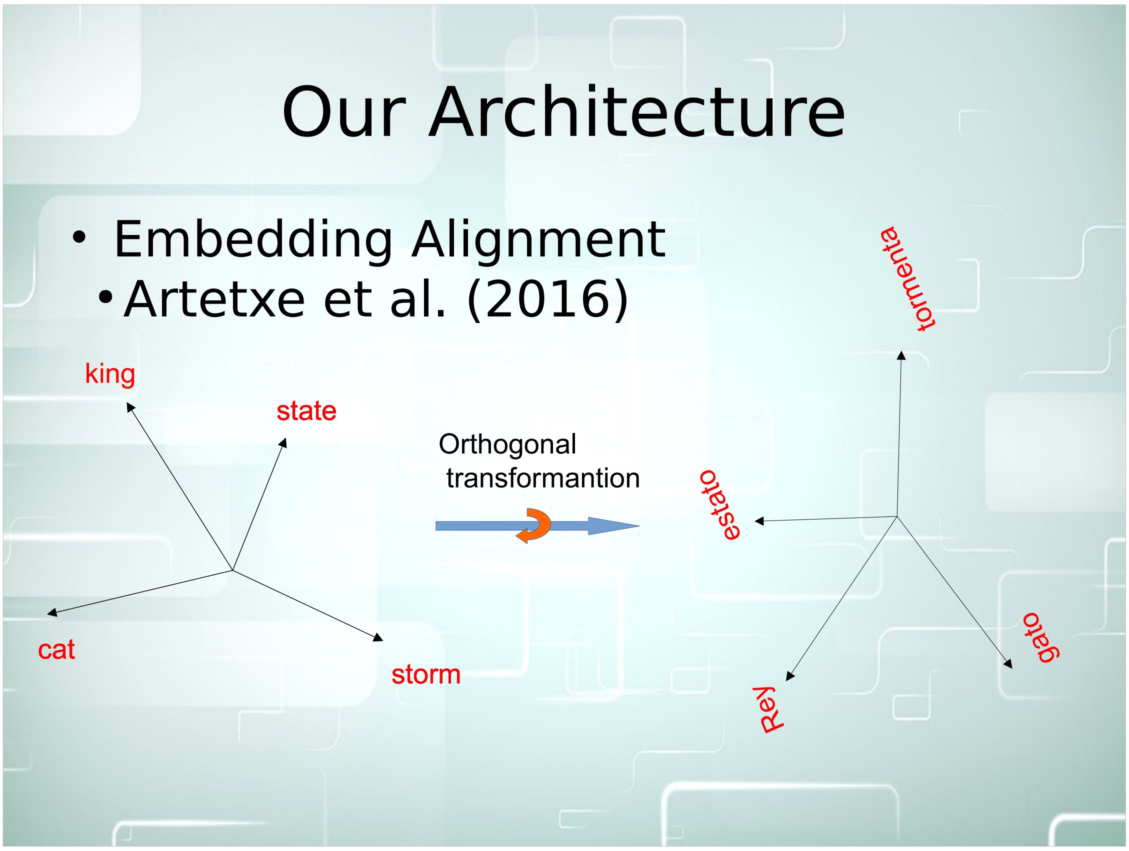
Third – Use the weighed sum to find the word that is most likely to be the answer

Last – recovery: Repr → ID

Now we talk about how we perform transfer learning.

The procedure of context matching and entity mapping are not inherently language dependent and can be shared

However the word vectors are always language-dependent

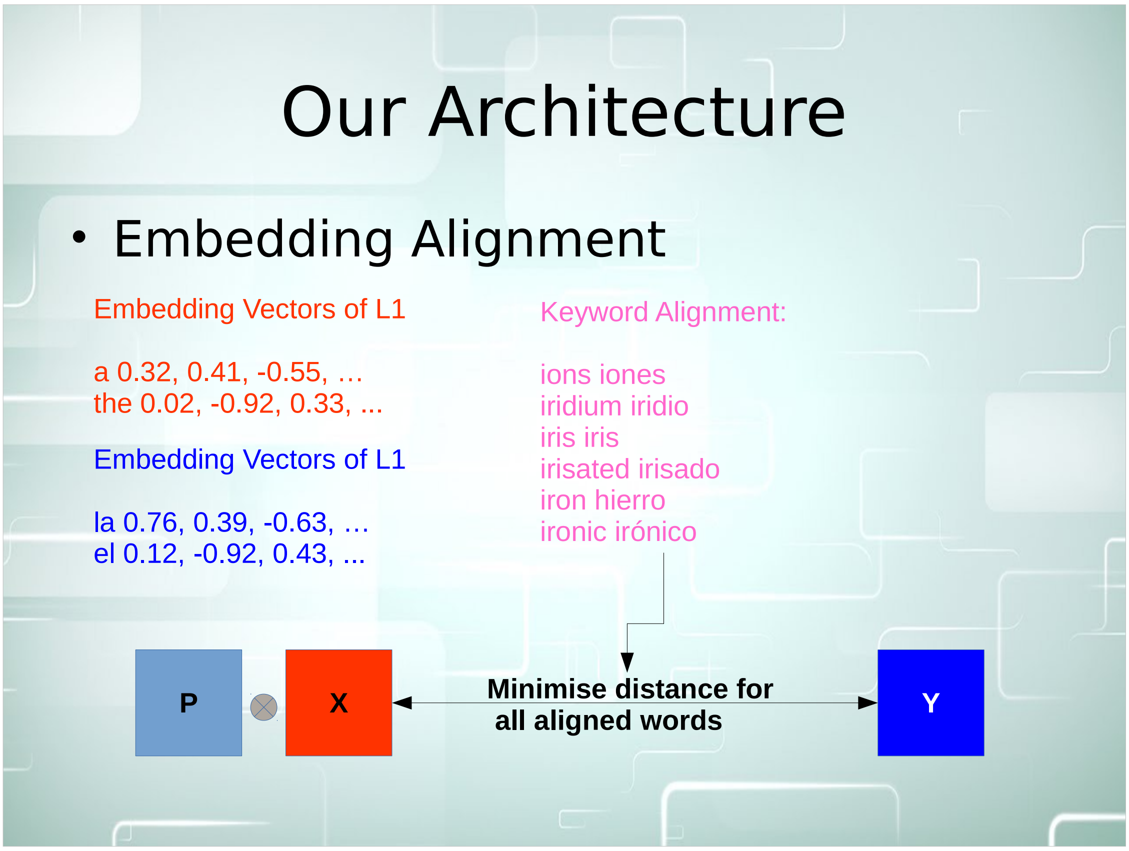


Fortunately, it is possible to map the word embeddings of two different languages into the same space such that the distance between similar words in two languages are mimimised

Word embeddings are learned from some form of collocation data and they encode syntactical and semantical information using the relative distance and position of word vectors

Assuming two languages are not too different and the corpus on which the embeddings are trained on are similar, they should have similar relative layout of equivalent words in the embedded space

Like two maps we can align them (via orthogonal tf)



Here we will talk about the alignment procedure. Suppose we have embedding vectors of Lang 1 and Lang 2, along with a list of words that are roughly equivalent in two aligned documents of these two languages

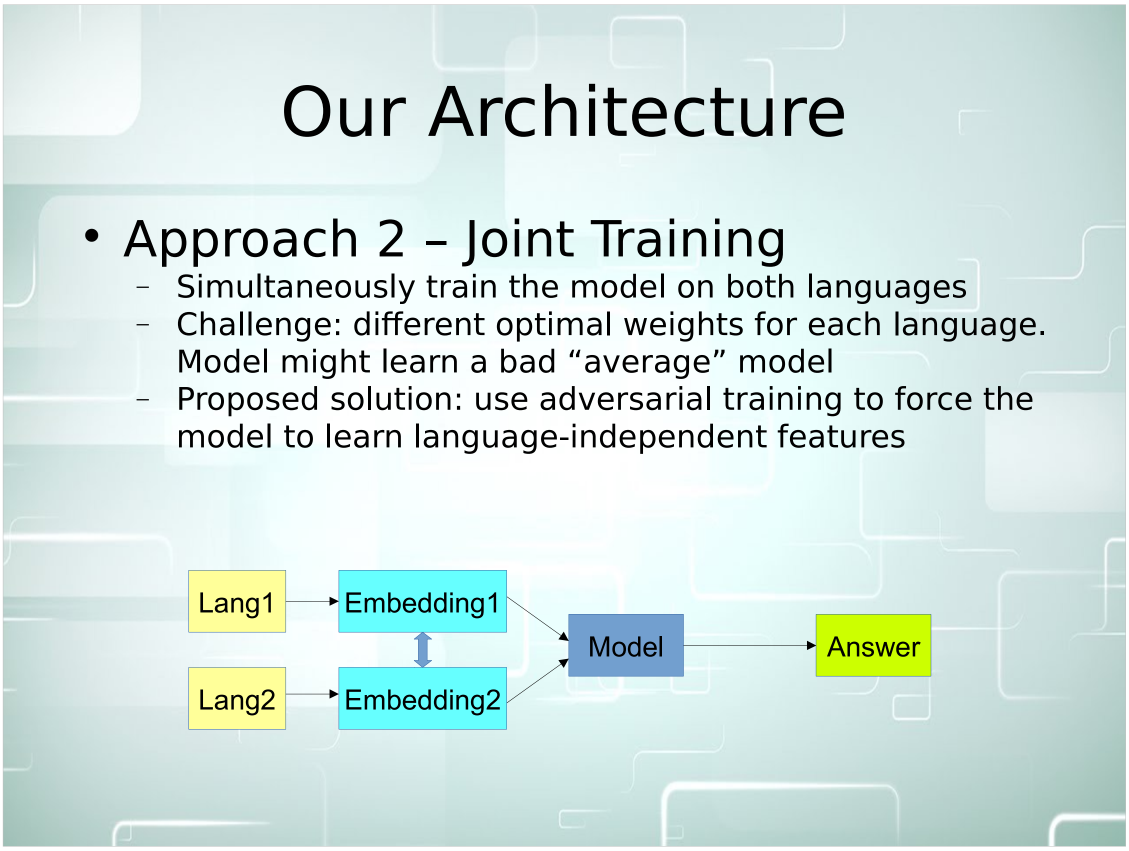
We find an optimal orthogonal transformation P such that the distance between aligned words in two languages are minimised

Strengths:

1. Similar words will generally be closer together
2. Does not impact monolingual information

Limitations:

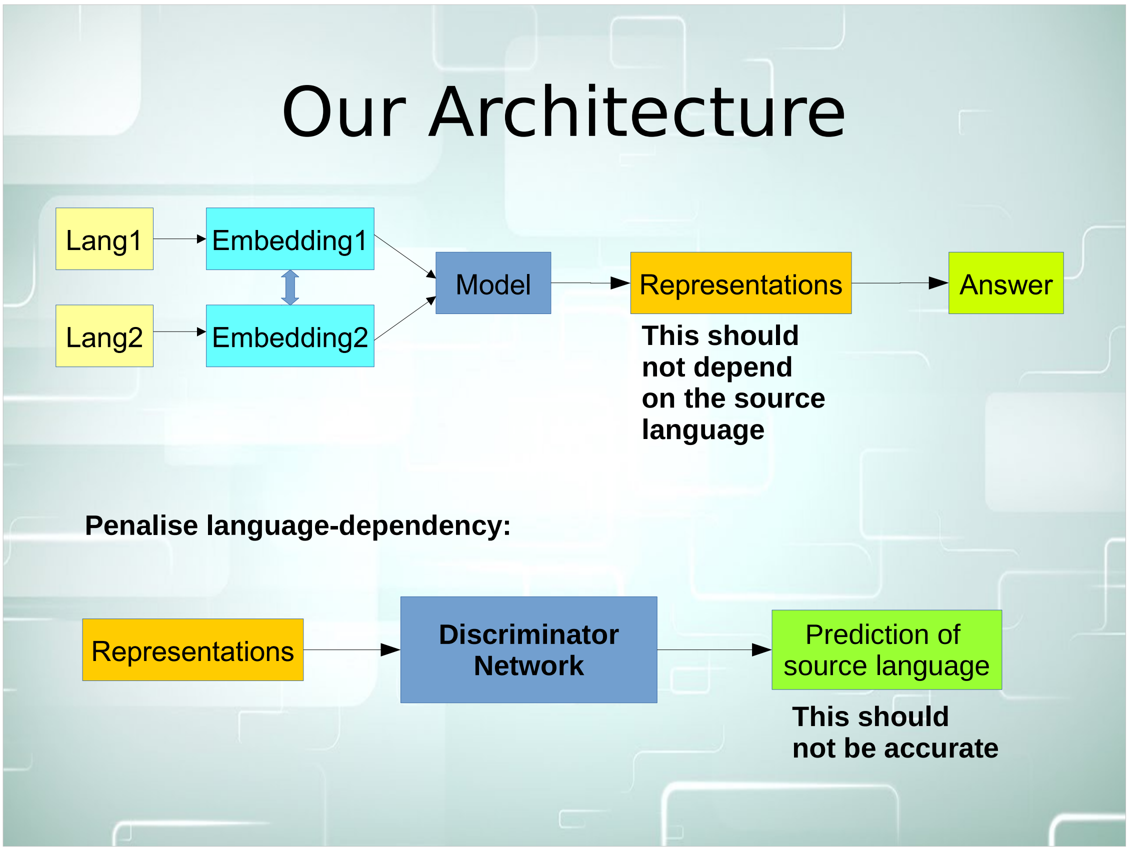
1. Alignment not perfect – multiple meanings, perfect equivalence might not exist
2. out-of-dictionary words are not guaranteed to be mapped close to similar words in the other language



Differece: instead of training sequentially, we train two models together, mixing the training examples

Difficulties:
1. “average” model
2. “overwhelmed” by majority language

Solution: adversarial training (discuss next)

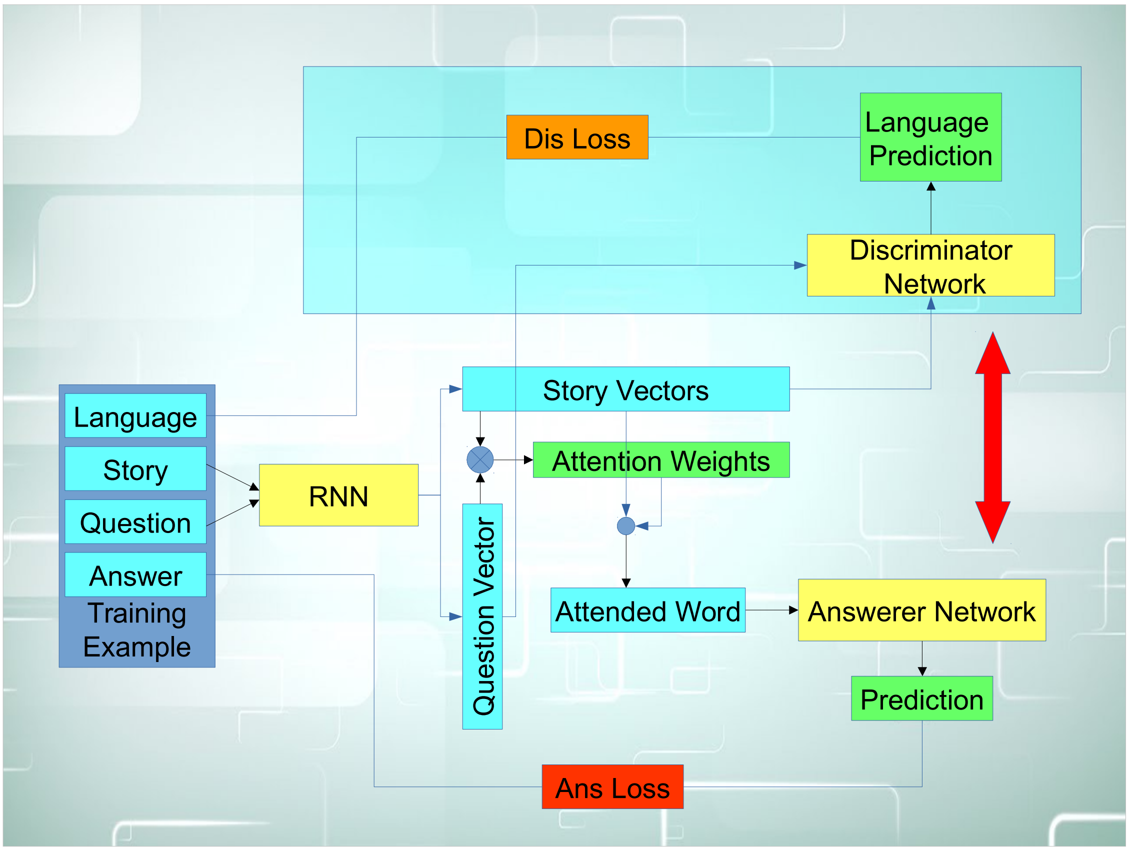


Model for two languages, from which a representation for the attended word is learned, which is used to predict the answer

We wish to learn a representation that does not depend on a specific source language and thus encodes language-agnostgic knowledge about the problem

We can achieve this by adding a penalty term to punish the network for learning language-dependent representations (but no such function exist out of the box)

But we have to train the penalty term ourselves through a discriminator network



Dis Loss
Language
Prediction
Discriminator
Network
Language
Story
Question
Answer
Training
Example
RNN
Story Vectors
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

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

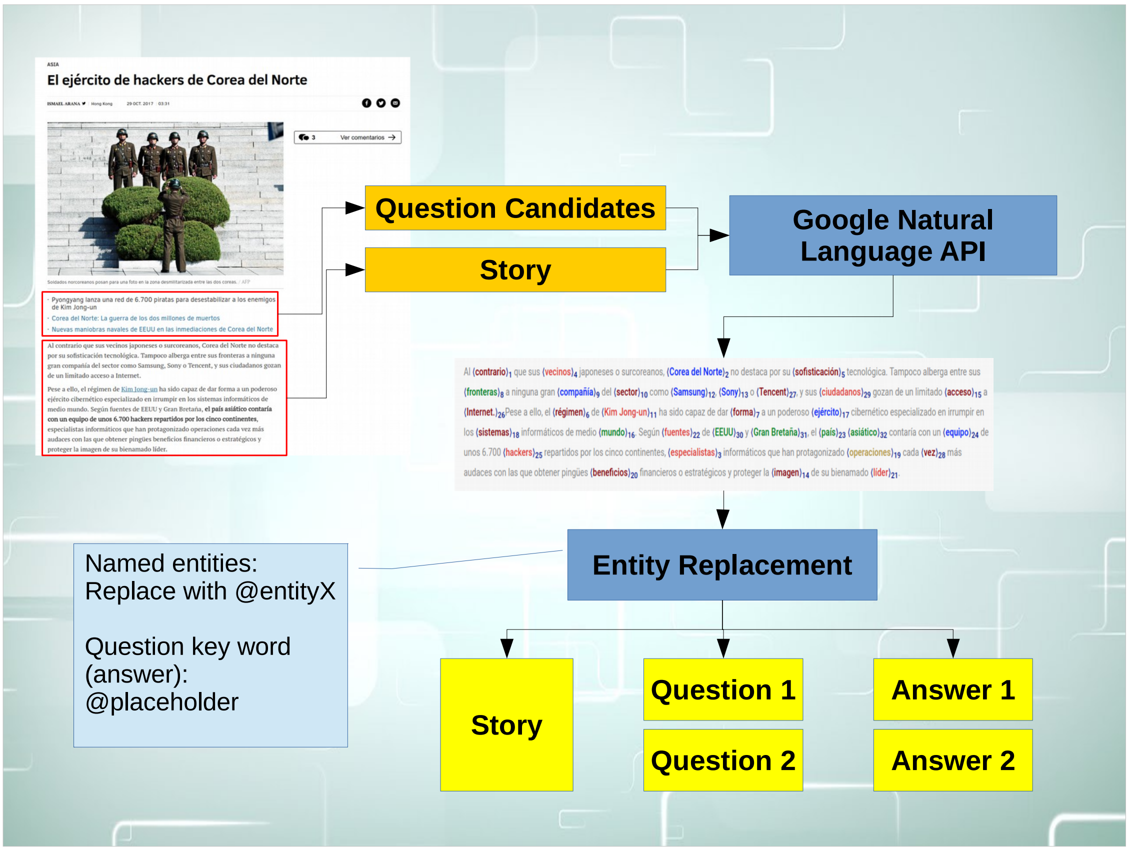
Follow the negative gradient of theta_D so that we minimise discrim. network loss wrt to it

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


El ejército de hackers de Corea del Norte
Question Candidates
Story
Google Natural Language API
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)

Points:

1. using mapped embeddings do not significantly degradate monolingual performance

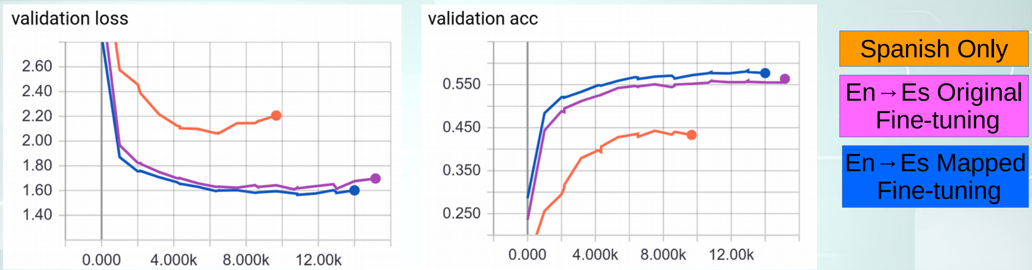
2. Best result on English monolingual performance is actually from using adversarial training (surprising)

3. Using fine-tuning transfer learning has significant beneficial impact on Spanish performance

4. Using mapped embeddings can further improve tranfer learning performance on Spanish data

5. Joint learning / Adversarial learning does not seem to perform better for Spanish data than fine-tuning + mapped embeddings

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