Computer Vision meets Natural Language Processing @ TrecVid 2016

Haithem Afli & Debasis Ganguly

Machine Learning Dublin Meet Up

November 28th, 2016

ADAPT : The Global Centre of Excellence for Digital Content and Media Innovation

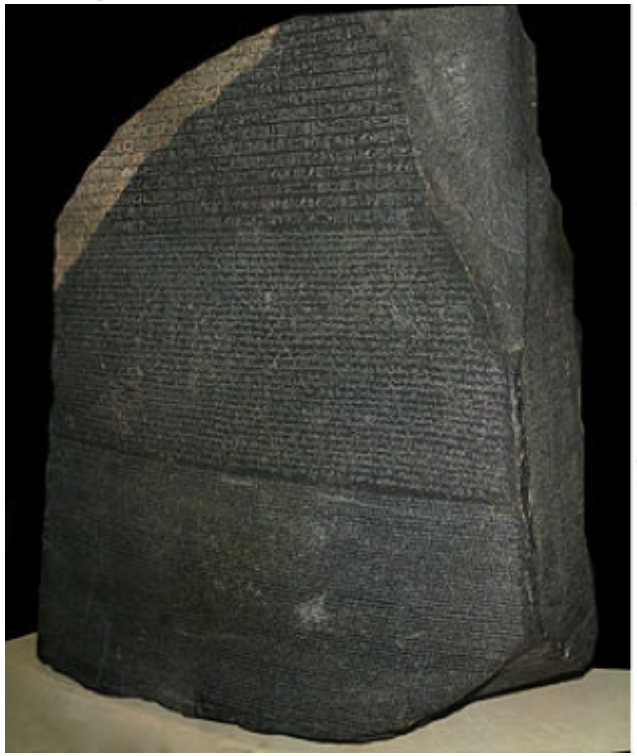
- Member of the ADAPT Machine Translation team led by **Prof. Andy Way**
- Manager of the ADAPT Social Media research group

IBM Dublin Research Lab

- Research Staff Member, IBM Research Lab, Dublin
- Former post-doctoral researcher, ADAPT Centre, DCU.

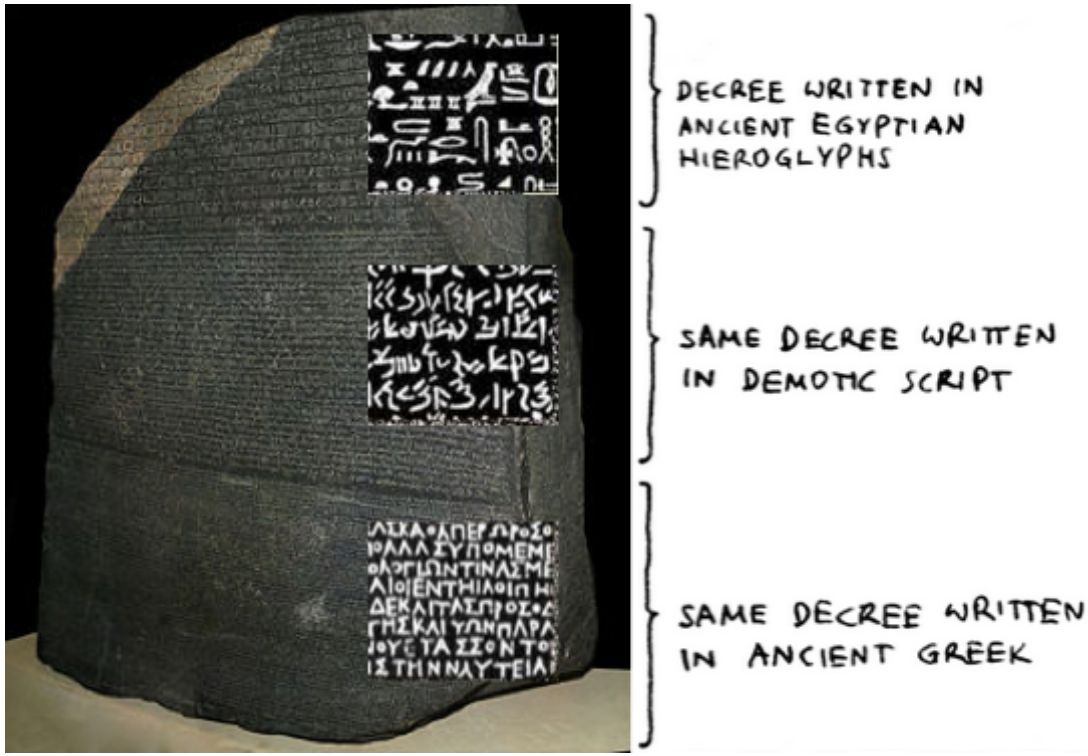
Natural Language : An age-old industry ?

- If you think the language industry is new
  $\rightarrow$ think again !



Natural Language : An age-old industry ?

- If you think the language industry is new
  $\rightarrow$ think again !



Rosetta Stone (British Museum)

- Carved in 196 BCE and re-discovered in 1799

Natural Language : An age-old industry ?

- For as far back as we can see, human has needed to communicate $\rightarrow$ so the origin of language industry is closely intertwined with the need of communication itself

The Tower of Babel and The House of Wisdom in Bagdad (Bait-al-Hikma)

- The work they produced paved the way for **the renaissance of culture !**

The age of social media

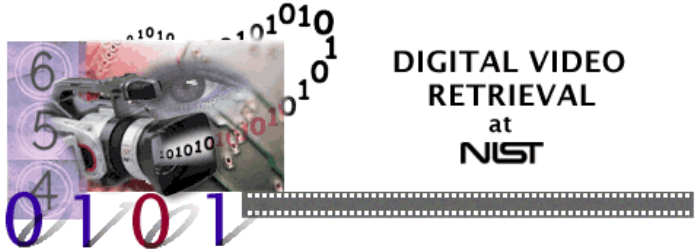
- Rapid growth of user-generated content available on the Web
  - Facebook updates, tweets on Twitter, WhatsApp messages, Youtube videos, etc.

$\rightarrow$ individual users have been able to actively participate in the generation of online content in different **modalities** (Text, images and vidoes)

$\Rightarrow$ caption generation models become a strong technique to capture and determine objects in the images and express their relationships in natural language.

TRECVID

- TREC Video Retrieval Evaluation (TRECVID) goal is to promote progress in content-based analysis of and retrieval from digital video
- In 2001 and 2002 the TREC series sponsored a video "track" devoted to research in automatic segmentation, indexing, and content-based retrieval of digital video
- Beginning in 2003, this track became an independent evaluation (TRECVID)



TREC Video Retrieval Evaluation: TRECVID

- Over the last 15 years TRECvid has had tasks in
  - shot bound detection, concept detection, instance search, known item search, example search, surveillance video event detection, multimedia event detection, video summarisation

→ New pilot on captioning

TREC Video Retrieval Evaluation: TRECVID

2016 Showcase Task - Video to Text Description

- Goals and Motivations
  - Measure how well can automatic system describe a video in natural language.
  - Measure how well can an automatic system match high-level textual descriptions to low-level computer vision features.
  - Transfer successful image captioning technology to the video domain.
- Real world Applications
  - Video summarization
  - Supporting search and browsing
  - Accessibility - video description to the blind
  - Video event prediction

IBM Research

Video Dataset

- Crawled 30k+ Twitter vine video URLs.
- Max video duration == 6 sec.
- A subset of 2,000 URLs randomly selected.
- Marc Ritter's TUC Chemnitz group supported manual annotations :
  - Each video annotated by 2 persons (A and B).
  - In total 4,000 textual descriptions (1 sentence each) were produced.
  - Annotation guidelines by NIST :
    - For each video, annotators were asked to combine 4 facets if applicable :
    - Who is the video describing (objects, persons, animals,. . . etc)
    - What are the objects and beings doing ? (actions, states, events,. . . etc)
    - Where (locale, site, place, geographic,...etc)
    - When such as time of day, season, ...etc

IBM Research

Samples of captions

A	B
a dog jumping onto a couch	a dog runs against a couch indoors at daytime
in the daytime, a driver let the steering wheel of car and slip on the slide above his car in the street	on a car on a street the driver climb out of his moving car and use the slide on cargo area of the car
an asian woman turns her head	an asian young woman is yelling at another one that poses to the camera
a woman sings outdoors	a woman walks through a floor at daytime
a person floating in a wind tunnel	a person dances in the air in a wind tunnel

IBM Research

Task 1 : Matching & Ranking

Person reading newspaper outdoors at daytime

Person playing golf outdoors in the field

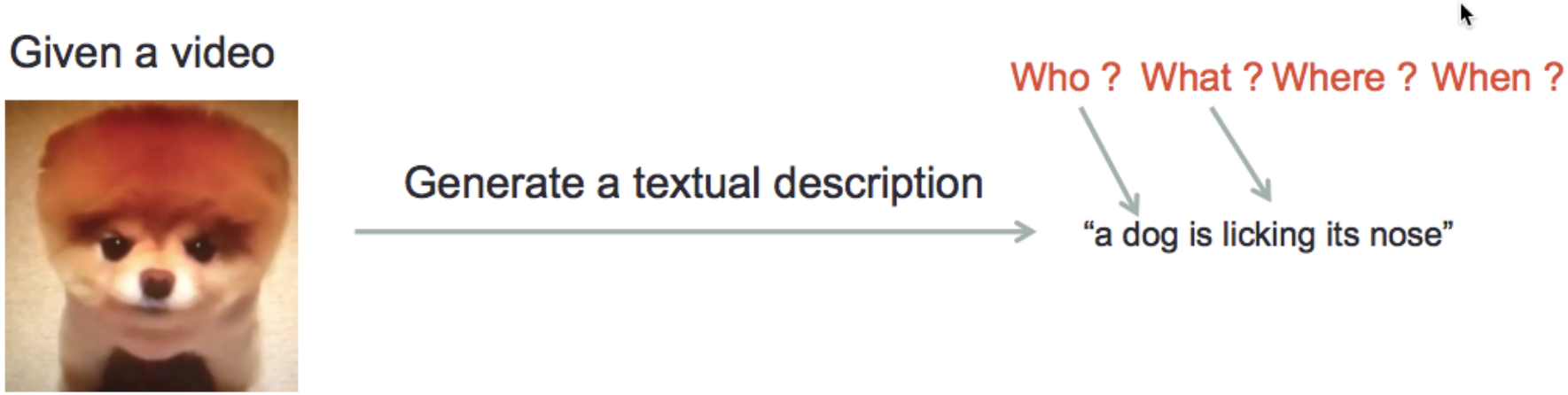
Three men running in the street at daytime

Two men looking at laptop in an office

x 2,000

x 2,000 type A … and … X 2,000 type B

Task 2 : Description Generation



- Metrics
  - Popular MT measures : BLEU , METEOR
  - Semantic similarity measure (STS).
  - All runs and GT were normalized (lowercase, punctuations, stop words, stemming) before evaluation by MT metrics (except STS)

Metrics

- **BLEU** [0..1] (bilingual evaluation understudy), used in MT to evaluate quality of text . . . approximate human judgement at a corpus level
  → Measures the fraction of N-grams (up to 4-gram) in common between source and target
- **METEOR** (Metric for Evaluation of Translation with Explicit Ordering) → Computes unigram precision and recall, extending exact word matches to include similar words based on WordNet synonyms and stemmed tokens
- **STS** measure [0..1] based on distributional similarity and Latent Semantic Analysis (LSA) . . . complemented with semantic relations extracted from WordNet

IBM Research

DCU participation

Engaging Content
Engaging People

Insight
Centre for Data Analytics

- Collaboration initiated by **Prof. Alan Smeaton**

IBM Research

DCU participation : Pre-Processing Vine Videos

- **Object Concepts :** We used the VGG-16 deep convolutional neural network to map keyframes in the videos to 1,000 object concept probabilities. We used 10 equally spaced keyframes per Vine video.
- **Behaviour Concepts** We applied crowd behaviour recognition to categorise the motion characteristics of a given Vine sequence. Keyframes are extracted fand probability scores calculated for 94 crowd behaviour concepts such as fight, run, mob, parade and protest.
- **Locations** were represented by extracting the probability scores from the softmax layer of VGG16 network pre-trained on the Places2 Dataset .

IBM Research

DCU participation : The Caption Generation Sub-Task

- we used an attention based model for automatic captions generation of images extracted from the VTT videos.
- Since we segmented the video into several static images, we generate one caption for each image of the video as one of the candidates for the video caption using **NeuralTalk2**, a CNN-RNN toolkit trained on the MSCOCO data set
- NeuralTalk2 takes an image and predicts its sentence description with a Recurrent Neural Network.

IBM Research

DCU participation : The Caption Generation Sub-Task

Example of attending to the correct caption

NeuralTalk2: a man is playing tennis on a tennis court

RefA: a man is playing tennis on a tennis court

RefB: a man plays tennis on a tennis court at daytime

Example of attending to the wrong caption

NeuralTalk2: a person holding a cell phone in their hand

RefA: 3 toy snowmen cry out in a room

RefB: three snowman plush toys wiggling on the floor at daytime

DCU participation : The Caption Ranking Sub-Task

- Caption matching task treated as an Information Retrieval (IR) task.
- IR : Given a query, retrieve a ranked list of documents sorted by similarity values.
- Query : Text comprised of the concept vector associated with each image.
- Retrievable document : The text associated with the captions.

IBM Research

DCU participation : The Caption Ranking Sub-Task

- Each concept vector is a fixed dimensional vector of 1000 dimensions.
- Query formulation strategy : Terms sorted by their component weights.
- Top k terms used for weighted query representation.
- BM25 used as retrieval model.

IBM Research

DCU participation : The Caption Ranking Sub-Task

Experiments Performed :

- Using different fields, i.e., places, objects, actions for query formulation.
- Aggregating (Averaging) the concept vector for each frame to the combined vector for the whole video.

Lesson Learned : Very good results on Caption Generation Ranking Tasks

NeuralTalk2: a man is playing tennis on a tennis court

RefA: a man is playing tennis on a tennis court

RefB: a man plays tennis on a tennis court at daytime

NeuralTalk2: a young boy playing soccer on a field

RefA: children play soccer on a soccer field

RefB: children plays soccer on a soccer field at daytime

NeuralTalk2: a dog and a cat laying on a couch

RefA: a cat sits on a bed

RefB: dog lies down on a bed

NeuralTalk2: a cat is sitting in a bathroom sink

RefA: a cat with a hat sits in a sink in a bathroom

RefB: cat with magician hat lying around in a bathroom sink

Continuation ..

- Motivated by our performance in caption ranking and caption generation we will refine our methods used in both tasks by broadening the number of underlying concepts.
- Continue the ADAPT + Insight collaboration with new partners such as IBM research Lab.

IBM Research

Many thanks for all the team members

Dublin City University and Partners' Participation in the INS and VTT Tracks at TRECVid 2016

Mark Marsden[1], Eva Mohedano[1], Kevin McGuinness[1], Andrea Calafell[3], Xavier Giró-i-Nieto[3] Noel E. O'Connor[1], Jiang Zhou[1], Lucas Azevedo[2], Tobias Daudert[2], Brian Davis[2], Manuela Hürlimann[2], Haithem Afli[4], Jinhua Du[4], Debasis Ganguly[4], Wei Li[4], Andy Way[4], Alan F. Smeaton[1]*

[1]Insight Centre for Data Analytics, Dublin City University, Dublin 9, Ireland
[2]Insight Centre for Data Analytics, National University of Ireland, Galway, Ireland
[3]Universitat Politècnica de Catalunya, Barcelona, Spain
[4]Adapt Centre for Digital Content Technology, Dublin City University, Dublin 9, Ireland

Abstract

Dublin City University participated with a consortium of colleagues from NUI Galway and Universitat Politècnica de Catalunya in two tasks in TRECVid 2016, Instance Search (INS) and Video to Text (VTT). For the INS task we developed a framework consisting of face detection and representation and place detection and representation, with a user annotation of top-ranked videos. For the VTT task we ran 1,000 concept detectors from the VGG-16 deep CNN on 10 keyframes per video and submitted 4 runs for caption re-ranking, based on BM25, Fusion, Word2Vec and a fusion of baseline BM25 and Word2Vec. With the same pre-processing for caption generation we used an open source image-to-caption CNN-RNN toolkit NeuralTalk2 to generate a caption for each keyframe and combine them.

1 Introduction

A team of researchers from Dublin City University (Insight and ADAPT Research Centres), National University of Ireland, Galway (Insight Research Centre) and Universitat Politècnica de Catalunya, took part in the TRECVid 2016 annual benchmarking [1], which is part of the TRECVid series [17] which first started in 2001. The team completed runs for two tasks namely Instance Search (INS) and the new showcase pilot task on Video to Text Description (VTT), both the matching and caption generation sub-tasks. These are described in this paper.

Thank you

DCU participation : Pre-Processing Vine Videos -More info

- **Object Concepts :** We used the VGG-16 deep convolutional neural network to map keyframes in the videos to 1,000 object concept probabilities. We used 10 equally spaced keyframes per Vine video. The model was pre-trained on the ImageNet ILSVRC training data, which consists of approx 1.3 million training images in 1,000 non-overlapping categories.
- **Behaviour Concepts** We applied crowd behaviour recognition to categorise the motion characteristics of a given Vine sequence. Keyframes are extracted fand probability scores calculated for 94 crowd behaviour concepts such as fight, run, mob, parade and protest. The mean concept score vector is then taken across the keyframes for a given Vine. These 94 concepts are taken from the WWW (Who What Where) crowd dataset which contains 10,000 video sequences fully annotated for all concepts

IBM Research

DCU participation : Pre-Processing Vine Videos -More info

- **Locations** Locations were represented by extracting the probability scores from the softmax layer of VGG16 network pre-trained on the Places2 Dataset [21]. This dataset contains over 1.8M images from 365 different scene categories (e.g. airport terminal, cafeteria, hospital room), which makes prediction of this network very suitable for this task.

IBM Research