Generating Headlines with Text Summarization

Bc. Dalibor Mészáros, supervisor: Dr. Márius Šajgalík

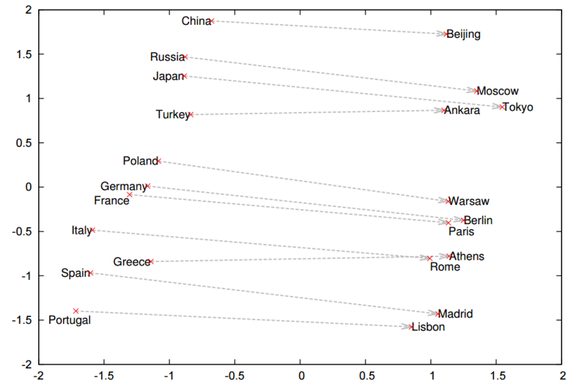
STU FIIT

Overview

- Exploring usage of **deep learning** in **text processing** field
- Idea based on adapting **sequential translation** and **summarization** approaches
- By using baseline summarization on raw text we can compress information to construct fitting **headlines**
- We evaluate our approach on datasets of Slovak and English Wikipedia articles and annotated documents

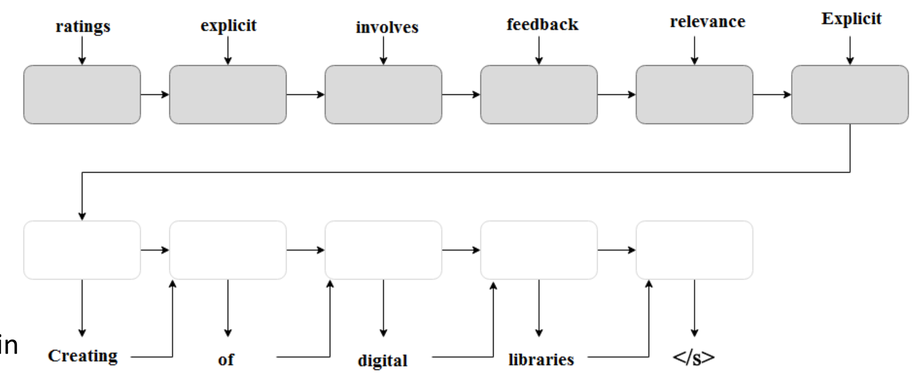
Distributed Representation of Words

- We represent words as **vectors** rather than dictionary indices
- Encoding the word's **positions** in vector space exposes **relationships** with other words
- We can compute such model using existing texts and learn the words **meaning** and **context** usage
- Training such model independently is relatively simple and further improves the cold start training of other dependent models such as **sequence translation models**



Sequence to Sequence Translation

- Popular solution for **machine translation** between two languages, because it doesn't map individual words between the domains, but rather sequences; preserves **context**, recognizes **multiword phrases** and the sequences may be **different in length**
- Words of both spaces are represented by initially random **embeddings**—using distributed representation of words
- Utilizing pre-trained **model of distributed representation**, we can further accelerate the training of sequential translation
- **The model consists of:**
- **Encoder**: which accepts the document's word embeddings in reversed order as they are positioned in sentences
- **Decoder**: which generates the expected headline's embeddings in correct order, ending with a stopping symbol. Each single generated word is further propagated during headline generation
- The model can output words, which are not included in input text simulating **abstractive** summarization



Results

- The model has it's own limitations as the **ratio** of input, output sequence and the **length** of sequences highly affects the **quality** of generated headlines.
- The model has **high memory requirements**

Metrics	Wikipedia SK	Annotated documents
Length of input	50	150
Length of output	5	20
Loss	2.7255	4.1221
Bleu 1-gram	73.91%	40.83%
Bleu 4-gram	30.08%	11.23%
Bleu-10 gram	-	02.34%
Meaningful words	44.75%	38.26%
Perfect prediction	21.09%	00.00%

- The model often yields **interesting** and **unexpected** outputs

Expected: An indoor navigation system for visually impaired and elderly people based on Radio Frq. **Predicted:** Autonomous navigation system with multiple level data
Expected: A Query Construction Service for Large Scale Web Search Engines **Predicted:** User behavior for personalized web
Expected: Content based image retrieval system using NOHIS tree **Predicted:** Improving image quality based on cluster information
Expected: Mužská štvorhra na WCT **Predicted:** Mužská štvorhra na Heineken
Expected: 1 slovenská národná futbalová **Predicted:** 1 slovenská hokejová liga