Bidirectional LSTM Recurrent Neural Network for Keyphrase Extraction

Marco Basaldella, Elisa Antolli, Giuseppe Serra, Carlo Tasso

Artificial Intelligence Laboratory @ University of Udine - http://ailab.uniud.it

Distiller improvements

- Two key elements:

Word Embedding

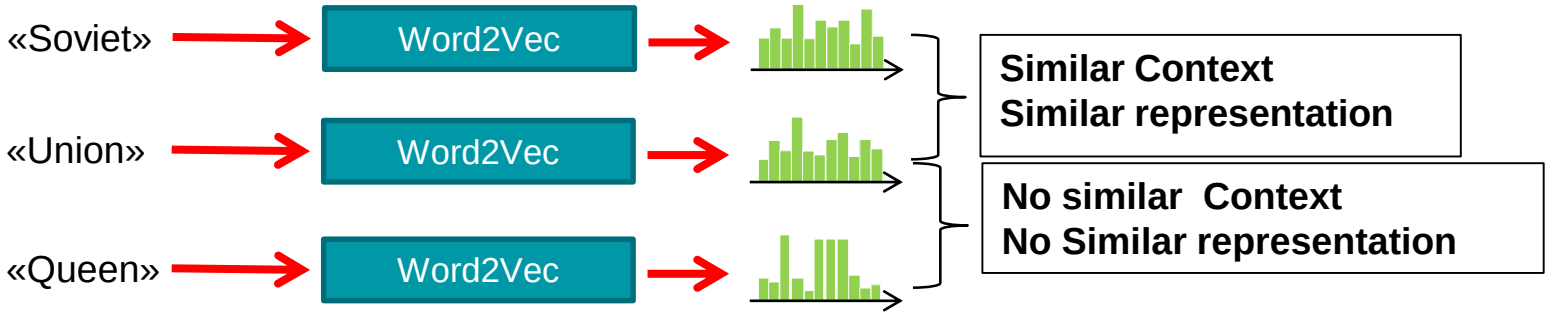
+

Recurrent Neural Networks

Word Embedding

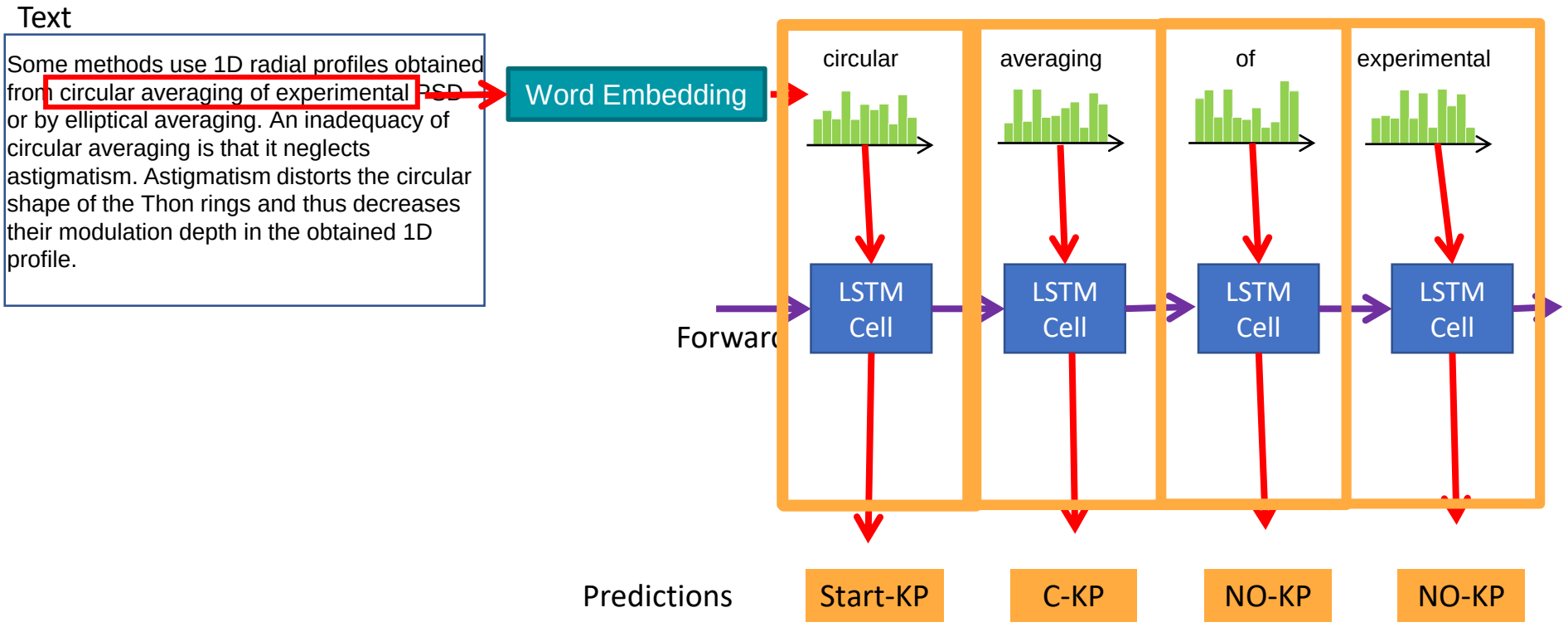
- **Word2Vec network** is a technique for building a rich semantic word embedding space (Google in 2013)
- **Key idea: two words have similar word embedding representations if they have a similar contexts**

- **For example:**



- We adopted Stanford's Glove Embeddings (trained with 6 billion words extracted from Wikipedia and Web texts)

Recurrent Neural Networks (LSTM)



Bidirectional Recurrent Neural Networks (BLSTM)

Future context is important in text understanding

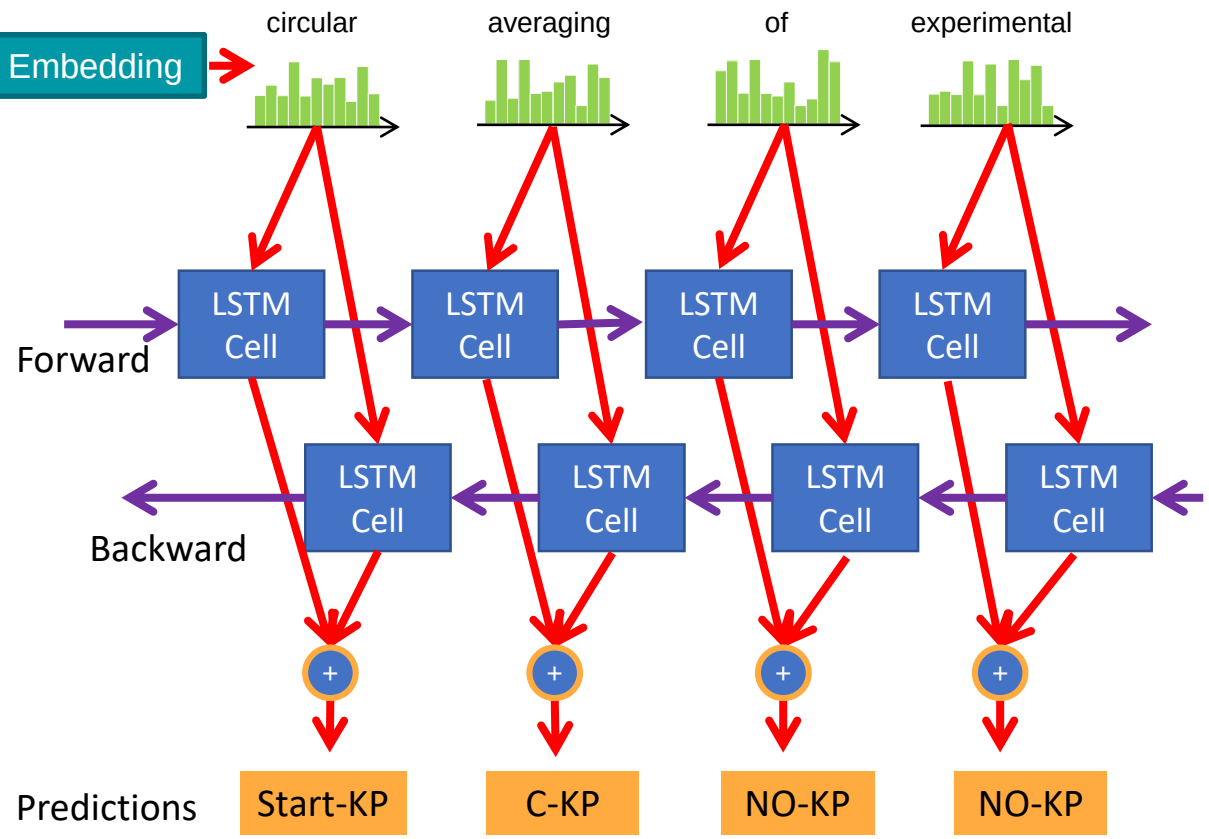
Toy example:

- «I speak very well **Japanese**, because I lived two years in Japan»

Our Full solution

Text

Some methods use 1D radial profiles obtained from circular averaging of experimental CSD or by elliptical averaging. An inadequacy of circular averaging is that it neglects astigmatism. Astigmatism distorts the circular shape of the Thon rings and thus decreases their modulation depth in the obtained 1D profile.



Experimental Results

- Inspec Dataset:
  - It’s one of the largest dataset for keyphrase extraction task
  - 2000 scientific abstract papers from Computers and Control e Information Technology
  - Experimental setting:
    - Training set: 1000 documents
    - Validation set: 500 documents
    - Test set: 500 documents
- Evaluation metrics:
  - Precision, Recall, F1-Score, Map, F1@5, F1@10

Word Embeddings evaluation

- **Test #1**: Training a Word Embedding Network on Inspec dataset does not perform well (Dataset too small)

- **Test #2:** We perform experiments using the pre-trained **Stanford's GloVe Embeddings**

Embedding	Size	Precision	Recall	F1-Score	Map	F1@5	F1@10
GloVe-50	50	0.297	0.637	0.405	0.336	0.271	0.333
GloVe-100	100	0.346	**0.653**	0.453	0.373	0.301	0.378
GloVe-200	**200**	**0.380**	0.642	**0.477**	**0.390**	**0.320**	**0.404**
GloVe-300	300	0.359	0.639	0.460	0.376	0.311	0.382

Comparison Results

- Comparision with "traditional" approaches:

Method	Precision	Recall	F1-Score
Proposed Approach	**0.380**	**0.642**	**0.477**
N-grams [1]	0.252	0.517	0.339
Noun Phrase Chunking [1]	0.297	0.372	0.330
Pattern [1]	0.217	0.399	0.281
Topic Rank [2]	0.348	0.404	0.352

- Comparision with a recent approach based on Deep Learning

Method	F1@5	F1@10
Proposed Approach	**0.320**	**0.404**
Deep KP Extraction [3]	0.278	0.342

[1] Hulth: Improved automatic keyword extraction given more linguistic knowledge. In: Proc. of Conference on Empirical Methods in Natural Language Processing (2003)
[2] Bougouin *et al.*: Graph-based topic ranking for keyphrase extraction. In: Proc. of Conference on Natural Language Processing (2013)
[3] . Meng *et al.* : Deep Keyphrase Generation. In: Proc. of Annual Meeting of the Association for Computational Linguistics (2017)

Some Examples #1

A simple **graphic approach** for **observer decomposition**
Based upon the proposition that the roles of inputs and outputs in a physical system and those in the corresponding **output-injection observer** do not really have to be consistent, a systematic procedure is developed in this work to properly divide a set of **sparse system models** and **measurement models** into a number of **independent subsets** with the help of a visual aid. Several smaller sub-observers can then be constructed accordingly to replace the original one.
The size of **each sub-observer** may be further reduced by strategically selecting one or more appended states. These techniques are shown to be quite effective in relieving **on-line computation** load of the **output-injection observers** and also in identifying **detectable sub-systems**.

Legenda

- Highlighted words: Results of the proposed approach
- **Underline bold**: Ground truth

Some Examples #2

BioOne: a new model for scholarly publishing
This article describes a unique electronic journal publishing project involving the **University of Kansas**, the **Big 12 Plus Libraries Consortium**, the **American Institute of Biological Sciences**, **Allen Press**, and **SPARC**, the **Scholarly Publishing** and **Academic Resources Coalition**. This partnership has created BioOne, a database of 40 full-text society journals in the biological and **environmental sciences**, which was launched in April, 2001. The genesis and development of the project is described and financial, technical, and **intellectual property models** for the project are discussed. **Collaborative strategies** for the project are described ….

Note: probably the «scholarly publishing» was selected, because in the training set there are several time the keyphrase «scholarly publishing model>

Legenda

- Highlighted words: Results of the proposed approach
- **Underline bold**: Ground truth