LaSTUS/TALN+INCO @ CL-SciSumm 2018 - Using Regression and Convolutions for Cross-document Semantic Linking and Summarization of Scholarly Literature

Ahmed AbuRa’ed, Àlex Bravo, Luis Chiruzzo, Horacio Saggion
Introduction

Provided a corpus and gold manual annotations indicating the facet and the text span(s) in the reference paper that best represent each citance.

Task 1A: For each citing sentence in the Citing Paper (citance), identify the spans of text (cited text spans) in the Reference Paper that most accurately reflect the citance.

Task 1B: For each cited text span, identify what facet of the paper it belongs to, from a predefined set of facets, namely: Aim, Hypothesis, Implication, Results or Method.

Task 2: generate a summary of the reference paper with a maximum length of 250 words.
Example

Reference paper (ACL ID: J96-3004) Citing paper (ACL ID: W12-1011)
Title: A Stochastic Finite-State Word-Segmentation Algorithm for Chinese
Title: Title: A Classical Chinese Corpus with Nested Part-of-Speech Tags
Task 1A: Unsupervised Methods

Reference doc

S1
S2
S3
S4
...

Citation doc

Distance metric

Vector space
Task 1A: Unsupervised Methods

Reference doc

S1
S2
S3
S4
...

Citation doc

Cit1

Distance metric

TF IDF
BabelNet embeddings centroid

Vector space

Modified Jaccard
Cosine similarity
## Task 1A: Unsupervised Methods

Performance over the development set (2017 Test Set)

<table>
<thead>
<tr>
<th>Method</th>
<th># Sents</th>
<th>Precision</th>
<th>Recall</th>
<th>F-Measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>MJ</td>
<td>1</td>
<td>0.120</td>
<td>0.103</td>
<td>0.111</td>
</tr>
<tr>
<td>MJ</td>
<td>2</td>
<td>0.072</td>
<td>0.115</td>
<td>0.088</td>
</tr>
<tr>
<td>BN</td>
<td>1</td>
<td>0.123</td>
<td>0.105</td>
<td><strong>0.113</strong></td>
</tr>
<tr>
<td>BN</td>
<td>2</td>
<td>0.085</td>
<td>0.142</td>
<td>0.106</td>
</tr>
</tbody>
</table>
Task 1A: CNN

Sentence level features

- **Position Features**
  - Sentence position
  - Sentence section position
  - Highest Probability Facet

- **Text Similarity Features**
  - Word Context Vector Cosine Similarity
  - Babelnet Synsets Vector Cosine Similarity

- **Dr Inventor Sentence Related Features**
  - Citation marker
  - Cause and effect
  - Co-reference Chains

- **Scientific Gazetteer Features**
  - Concept Lexicon
  - Action Lexicon
Training

Regression Problem.

CNNs with 3 windows to represent bi, tri and quad-grams.

22 features based on the pair of sentences (reference and citing).

Scoring based on the distance between the sentence and the nearest cited sentence.
Task 1A: Voting system

Combine previous methods:

- Top 10 CNN sentences
- Top 10 MJ sentences
- Top 10 BN sentence
- Top 40 Task 2 sentences

Take the intersection of results

- Sentences 2 systems agree on
- Sentences 3 systems agree on
Task 1B

Same neural network as Task 1A

Classification Problem

- Aim
- Implication
- Method
- Results
- Hypothesis

Last layer softmax.
Task 2: Text Summarization

Neural Network based on Convolutions

- The network generates a summary by selecting the most relevant sentences from the RP.

- The CNN is fed by a combination of linguistic and semantic features from RP and CPs → **Context Features** and **Word Embeddings**.

- The aim of our CNN is to learn the **relation between a sentence and a scoring value** indicating its relevance.
Task 2: Text Summarization

Context Features

- **Linguistic information** extracted from the cluster (RP and CPs)

- This information is based on a set of **numeric features** to characterize each sentence in the RP and its relation with the corresponding CPs.

- Before the extraction of context features:
  - *Computation of Word Vectors* based on Word Embeddings (Google and ACL) and frequencies (SUMMA).
  - For each sentence (RP and CPs) → 3 sentence vectors based on centroid.
  - Only for the RP, we also compute the 3 vectors related to the abstract and the whole article.
Task 2: Text Summarization

Context Features

- From the sentence vector representations, we extracted a set of numeric features some of which are based on comparing a sentence to its (document or cluster) context:
  - Sentence Abstract Similarity Scores
  - Sentence Centroid Similarity Scores
  - Sentence Title Similarity Scores
  - Position in the Document Score
  - Position in the Document Section Score
  - TextRank Normalized Scores
  - Term Frequency Score
  - Citation Marker Score
  - Rhetorical Class Probability Scores
  - Citing Paper Maximum Similarity Scores
  - Citing Paper Minimum Similarity Scores
  - Citing Paper Average Similarity Scores
Task 2: Text Summarization

Scoring Values for Learning

- CNN learns the relation between **features** (Word Embeddings and Context Features) and a **score value**.
- A **regression task** by devising various scoring functions to represent the likelihood of a sentence belonging to each gold standard summary.
- **Scoring functions**:
  - **Cosine Distance**: maximum cosine similarity between each sentence vector in the RP with each vector in the gold standard summaries.
  - **ROUGE-2 Similarity**: each sentence in the RP is compared with each gold standard summary using ROUGE-2.
  - **Scoring Functions Average**: we computed the average between all scoring functions. We also calculated a simplified average with vectors do not based on word-frequencies (no SUMMA).
Task 2: Text Summarization

CNN Model

Word Embeddings
Google + ACL

Context Features

Sent(i)
Sent(i-1)
Sent(i-2)
Sent(i-3)
Word(0)
Word(1)
Word(2)
Word(n)
Word(n-1)
Word(n-2)
Word(n-3)
...
Conclusions

For Task 1

- Neural network based on convolutions.
- The input is a combination of word embeddings with a set of features based on a pair of sentences (reference and citation sentences).
- The approach outperformed results reported in last year CL-SciSumm-17 Shared Task 1A.
- We believe that one system is not enough to identify which exact sentences have been cited in a reference paper.
Conclusions

For Task 2:

- Neural network based on convolutions.
- CNN uses a combination of word embeddings with sentence relevance and citation features to learn a specific Scoring Function.
- The approach outperformed results reported in last year CL-SciSumm-17 Shared Task 2.
Thanks!