Text Assisted Insight Ranking Using Context-Aware Memory Network

Qi Zeng*[1], Liangchen Luo*[2], Wenhao Huang[3], Yang Tang[2]

Stony Brook University[1], Peking University[2], Shanghai Discovering Investment[3] (* *indicates equal contribution*)

INTRODUCTION

Task: Automatically extract useful insights, i.e. the data mining results, from a multi-dimensional table, model their usefulness or interestingness, and rank the top ones.

Challenge: lack of available paired data; multi-dimensional comparison criteria.

Solution:

- Text Assistance
- Ranking Model
- Global Context-Aware Memory

	2017	2016	2015
Total Revenue	**2443299**	**2529619**	2218032
Cost of Revenue	861242	932240	729256
Gross Profit	1582057	1597379	1488776
Selling General and Administrative	1001307	1251105	1132164
Operating Income or Loss	38740	-367208	-450036
Interest Expense	**105237**	**99968**	98178

Total revenue was $2.44 billion, a decrease of 3% compared to 2016. In 2017, **interest expense** increased by $5.3 million compared to 2016.

CONTRIBUTIONS

- We formally formulate the problem of text assisted insight ranking, which has not been fully investigated yet.
- We construct a new financial dataset, in which insight importance is labeled with text assistance.
- We propose a context-aware memory network to model the importance of insights. The experimental results on two datasets show that our approach significantly outperforms the baseline methods.

MODEL

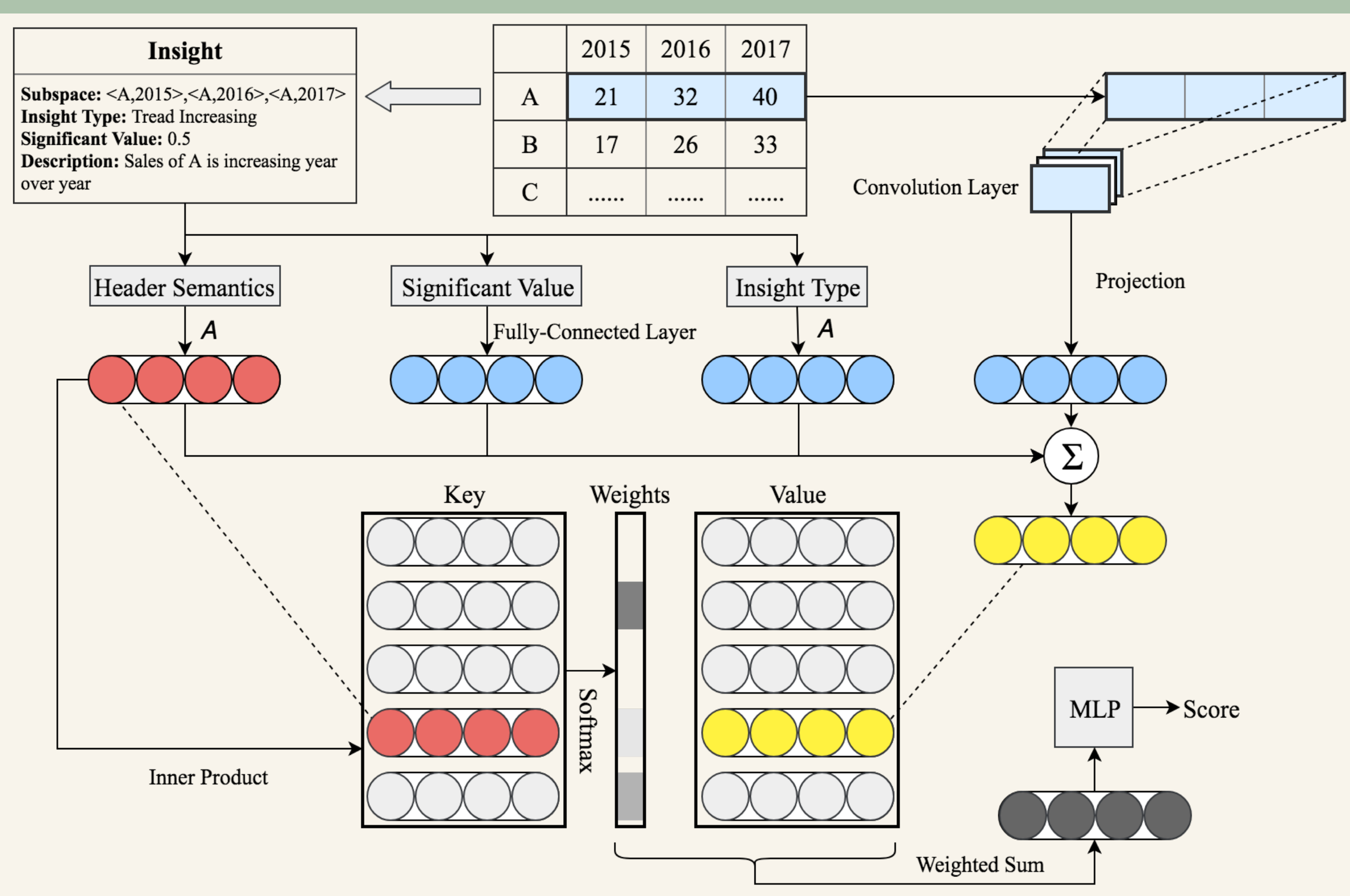
Neural Ranking module

The neural ranking model explores the data characteristics, including its semantic information, insight type, statistic information and subspace, and assigns importance scores to each insight.

The model is implemented as a multi-layer perceptron (MLP) and trained by minimizing the L2 loss $J(\gamma)$ of the output scores and the similarity scores of the insights.

Key-Value Memory Network

Additionally, the key-value memory network model introduces other insights within one group, namely the table context, into the ranking process.



EXPERIMENT

We conduct experiments on Financial Report Dataset and SBNation Dataset (Wiseman et al., 2017).

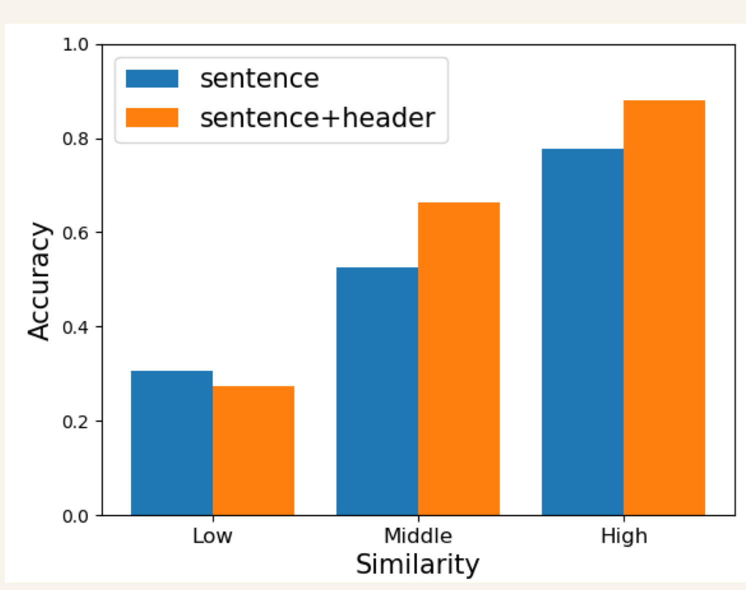
Financial Report Dataset Statistics

The dataset contains in total 5,670 reports and 49,129 tables of 2,762 companies from United States Securities and Exchange Commission.

	Mean	Percentile	
		5%	95%
# tokens per cell	5.29	1	12
# tokens per sentence	32.36	15	64
# sentences per report	774.98	282	1434

Accuracy of Text Assistance Method

We randomly sample 4,000 pairs of insights and their most similar sentences in the reports, and ask 10 annotators to label whether the pairs are of the same meaning.



Evaluation Results on SBNation Dataset

Models	Precision@1	Precision@3	Precision@5
$Sig_{cluster}$	0.727	0.629	0.540
TAR_{memory}	**0.886**	**0.813**	**0.745**

Evaluation Results on Financial Report Dataset

Models	Precision@1	Precision@3	Precision@5	mAP@3	mAP@5	NDCG@3	NDCG@5
Sig_{table}	0.098	0.246	0.399	0.474	0.624	0.646	0.688
$Sig_{dataset}$	0.107	0.249	0.408	0.473	0.621	0.649	0.692
$Sig_{cluster}$	**0.110**	**0.261**	**0.416**	**0.481**	**0.632**	**0.658**	**0.703**
TAR_{cnn}	0.118	0.278	0.444	0.525	0.686	0.738	0.757
$TAR_{semantics}$	0.162	0.411	0.605	0.668	0.756	0.799	0.815
TAR_{memory}	**0.170**	**0.425**	**0.626**	**0.684**	**0.772**	**0.812**	**0.829**

Case Study

It consists of the top 5 insights in 10 insight candidates from one table. The reason why the fourth insight is wrongly labeled is that the similarity score is incorrectly calculated and the gold standard is in fact inaccurate. This serves as an example of the optimization direction.

Insight Descriptions	Gold	TAR
Collaboration and license revenue was 71.7 million for the year ended, an increase of 58.7 million compared to the year ended.	1	2
General and administrative expenses were 27.8 million for the year ended, an increase of 18.8 million compared to the year ended.	2	4
Research and development expenses were 3 1 58.6 million for the year ended, an increase of 35.1 million compared to the year ended.	3	1
We had 111 full-time employees including 82 employees engaged in development	4	9
The net valuation allowance increased by 4.9 million and 0.6 million respectively.	5	3

MISC

References

1. Wiseman, S.; Shieber, S. M.; and Rush, A. M. 2017. Challenges in data-to- document generation. In *Proc. of EMNLP*.

Links:

arxiv.org/abs/1811.05563

Scan QR code to view the paper