

# Learning Thematic Similarity Metric Using Triplet Networks

Liat Ein-Dor, Yosi Mass, Alon Halfon, Elad Venezian, Ilya Shnayderman, Ranit Aharonov, Noam Slonim IBM Research AI

{liate,yosimass,alonhal,eladv,ilyashn,ranita,noams}@il.ibm.com

### **Motivation**

- Argumentation mining (Project Debater)
  - Organize arguments into themes to generate a compelling argumentative narrative
- Multi-document summarization

### Learn Thematic Similarity (DNN Triplet Network)



- Organize selected sentences into meaningful sections and paragraphs
- General sentence representations do not capture **thematic** similarity

### **Example - Trans fats usage in food should be banned**

"A new study has found that the ingestion of trans fats increase the risk of suffering depression."

"A study found that junk food high in trans fats can damage sperm in otherwise healthy, young men."

"Trans fats can increase the physical brain damage and manic behavior associated with amphetamine abuse."

"...if trans fat made up too much of an infant's intake of fat, it may affect brain and eye development."

"Recent research from the FDA has found that trans fats are linked to infertility in women."



### **Results on the clustering task**

### **Clustering benchmark**



Clustering (iclust)



# "Trans fats can have a bad influence on children's health."

## Approach

- Cluster sentences into themes
  - Requires a similarity metric between sentences dedicated to thematic clustering
- Train a Deep Neural Network to learn a thematic similarity through a comparative task.
  - Requires a large set of labeled data
- Create a large weak-supervised labeled data from the partition of Wikipedia articles into sections
- Create a clustering benchmark for evaluation



MI	AMI	RI	ARI
skip-thoughts-CS	mean-vectors		triplets-sen-rand-art
triplets-titles	triplets-sen	ets-sen <b>I</b> tripletsd-sen-titles	

### **Results on the triplets dataset**

Method	Accuracy
mean-vectors	0.65
skip-thoughts-CS	0.615
skip-thoughts-SICK	0.547
triplets-sen	0.74

### **Results on the SICK semantic relatedness task**

Method	Pearson	Spearman	MSE
mean-vectors	0.757	0.673	0.455
skip-thoughts-CS	0.858	0.791	0.287
triplets-sen	0.797	0.704	0.372
triplets-titles	0.786	0.685	0.393
triplets-sen-titles	0.818	0.724	0.339

### **Generation of Weakly-Labeled Triplets**

Wikipedia Articles



triplets-sen: Two sentences from the same section are closer than a third sentence from a different section

ttriplets-titles: A sentence is closer to its section's title than to another section's title

	# Triplets - Train/Validation/Test
riplets-sen	1.78M/220K/223K
riplets-titles	1.38M/172K/173K

### **Conclusion:**

Triplet networks, trained on weak-supervised data, outperform other methods on thematic clustering tasks and perform on par with them on other semantic relatedness tasks.

Get the clustering benchmark here

www.research.ibm.com/haifa/dept/vst/debating\_data.shtml



Project Debate