Computing Semantic Relatedness Using a Word-Text Mutual Guidance Model

Bingquan Liu, Jian Feng, Ming Liu, Feng Liu, Xiaolong Wang, and Peng Li

Harbin Institute of Technology

Dec. 8th, 2014

Content Outline

- Introduction
- Method
- Experiments
- Conclusion

Introduction

- The computation of semantic relatedness requires the estimation of the degree of association between two text fragments.
  - surface meaning, contextual knowledge
  - statistical information, semantic information
- Relatedness between two texts or words has great meaning.
  - Question match
  - Text clustering
  - Semantic computing

Related Work

- Knowledge-based methods.
  - Employ information extracted from manually constructed lexical taxonomies, e.g. WordNet
  - Focused on developing appropriate measures while using WordNet as the primary knowledge source and obtained relatively good results
- Corpus-based measures.
  - Employ probabilistic approaches to compute the semantic relatedness among words and texts
  - Map words or texts to the corresponding article in Wikipedia, which has emerged as a promising conceptual network for semantic relatedness
- The limit of existing methods.
  - Ignore the internal relationships between words and texts

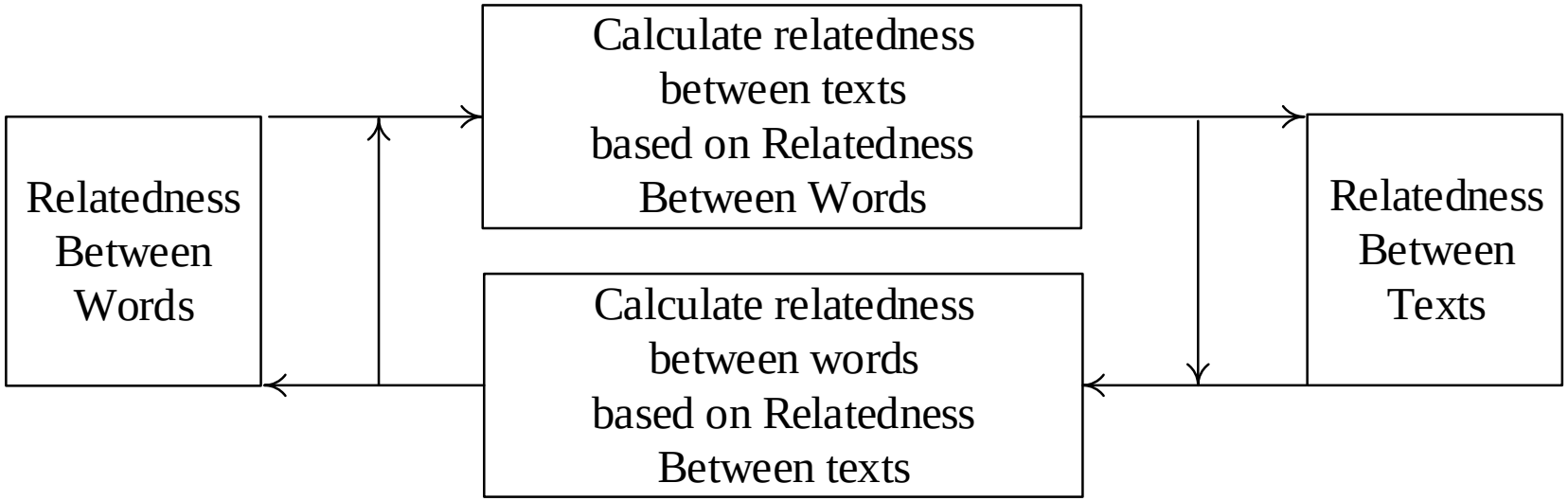
Our work

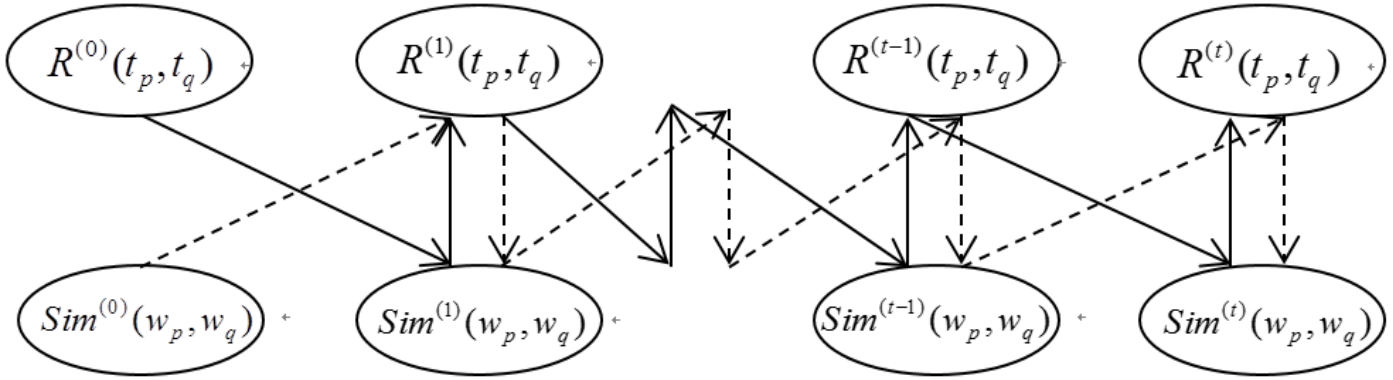
- WTMG, word-text mutual guidance model
- The mutual guidance between concept A and concept B is defined as a process where A can be derived from B and B can be derived from A.
- In our model, we compute the relatedness among words using the text relatedness and the relatedness among texts can be calculated based on the word relatedness.
- We propose an iterative process that computes the relatedness among words and texts.
- Two main steps:
  - First, we establish the initial word relatedness or text relatedness and we construct a relatedness matrix.
  - Second, the word relatedness and text relatedness are calculated iteratively.

Main contributions

- First, we propose the WTMG model to make full use of the internal relationships between words and texts.
- Second, we performed comparisons of many word and text semantic relatedness initialization methods.

Method - Global View





Method - Initialize Words

- Path-based measures
  - These measures compute the word relatedness as a function of the number of edges in the taxonomy along the path between two conceptual nodes c_1 and c_2 onto which the words w_1 and w_2 are mapped.
    - L&C method

$$sim_{L\&C} = -\log \frac{length(c_1, c_2)}{2D}$$

    - Wup method

$$sim_{Wup} = \frac{depth(LCS(c_1, c_2))}{depth(c_1) + depth(c_2)}$$

Method - Initialize Words

- Information content-based measures
  - Semantic relatedness between concepts are then calculated based on the information content.
    - The information content is defined as $IC(c) = -\log P(c)$

    Where $P(c)$ is the probability that a randomly selected word in a corpus is an instance of concept c.
    - Lin method

      $$sim_{Lin} = \frac{2IC(LCS(c_1, c_2))}{IC(c_1) + IC(c_2)}$$
    - J&C method

      $$sim_{J\&C} = \frac{1}{IC(c_1) + IC(c_2) - 2IC(LCS(c_1, c_2))}$$

Method - Initialize Texts

- A method for measuring the semantic relatedness of texts by exploiting the information that can be extracted from the relatedness of their component words.

$$sim(t_1, t_2) = \frac{1}{2}\left[\frac{\sum_{w \in \{t_1\}} \left(\max\left(w, t_2\right) \cdot idf\left(w\right)\right)}{\sum_{w \in \{t_1\}} idf\left(w\right)} + \frac{\sum_{w \in \{t_2\}} \left(\max\left(w, t_1\right) \cdot idf\left(w\right)\right)}{\sum_{w \in \{t_2\}} idf\left(w\right)}\right]$$

Method - Iteration

- Iterative procedure for calculating Sim(w_p, w_q)

$$sim^{(t+1)}\left(w_p, w_q\right) = \left(1-\lambda\right) sim^{(t)}\left(w_p, w_q\right) + \lambda \sum_{k=1}^{M}\left[\left(\sum_{i=1}^{M} tf_{ip} Q_{ik}^{(t)}\right)\left(\sum_{i=1}^{M} tf_{iq} Q_{ik}^{(t)}\right)\right]$$

- Iterative procedure for calculating R(t_p, t_q)

$$R^{(t+1)}\left(t_p, t_q\right) = \left(1-\lambda\right) R^{(t)}\left(t_p, t_q\right) + \lambda \sum_{k=1}^{N}\left[\left(\sum_{j=1}^{N} tf_{pj} P_{jk}^{(t+1)}\right)\left(\sum_{j=1}^{N} tf_{qj} P_{jk}^{(t+1)}\right)\right]$$

$$P_{jk} = \frac{sim\left(w_j, w_k\right)}{\sqrt{\sum_{l=1}^{N} sim\left(w_j, w_l\right)^2}} \qquad Q_{ik} = \frac{sim\left(t_i, t_k\right)}{\sqrt{\sum_{l=1}^{M} sim\left(t_i, t_l\right)^2}}$$

Experiments - evaluation

- Pearson's correlation coefficient

$$\gamma = \frac{\operatorname{cov}(X,Y)}{\sigma_X \sigma_Y} = \frac{E(XY) - E(X)E(Y)}{\sqrt{E(X^2) - E^2(X)}\sqrt{E(Y^2) - E^2(Y)}}$$

- Spearman's rank correlation coefficient

$$\rho = 1 - \frac{6\sum (x_i - y_i)^2}{n(n^2 - 1)}$$

Effectiveness of WTMG

- We selected five similar text pairs (designated as S1 to S5) and five dissimilar text pairs (designated as U1 to U5).
- We applied our WTMG model to this small corpus example and the results are shown in Figures 2 and 3.

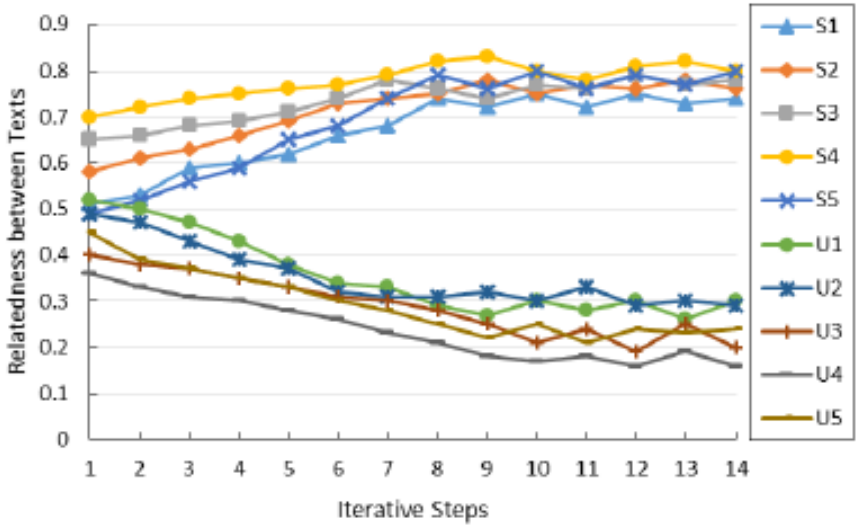


Fig. 2. Initialize Text Relatedness

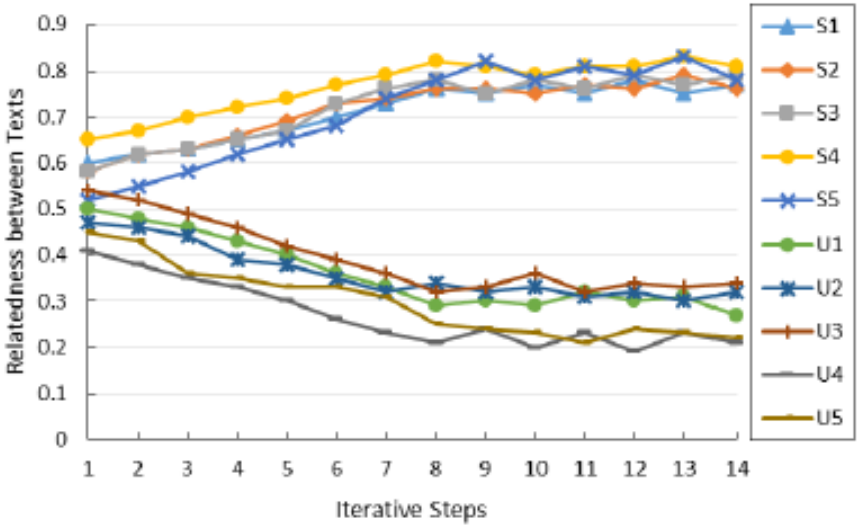


Fig. 3. Initialize Word Relatedness

Experiments - Result

- DataSet: Rubenstein and Goodenough(RG65), WordSimilarity-353(WS353), Mturk-771(MT771), Miller-Charles(MC30)

Pearson and Spearman results for the word relatedness datasets

Method		Pearson(γ)				Spearman(ρ)			
		MC30	RG65	WS353	MT771	MC30	RG65	WS353	MT771
Know-ledge-based	Wup	0.778	0.784	0.282	0.477	0.750	0.755	0.339	0.398
	J&C	0.695	0.731	0.354	0.498	0.820	0.804	0.318	0.402
	L&C	0.779	0.839	0.313	0.503	0.768	0.797	0.302	0.410
	Lin	0.835	0.858	0.329	0.513	0.750	0.788	0.348	0.424
	Resnik	0.813	0.836	0.362	0.431	0.693	0.731	0.353	0.404
Corpus-based	LSA	0.725	0.644	0.563	-	0.662	0.609	0.581	-
	ESA	0.588	-	0.503	-	0.727	-	0.629	-
	SSA	0.778	0.850	0.590	-	0.843	0.800	0.537	-
Ours	WTMGW	**0.879**	**0.861**	**0.622**	**0.572**	**0.846**	0.826	**0.750**	**0.480**
	WTMGR	0.871	0.847	0.602	0.539	0.820	**0.830**	0.748	0.477

The table shows that the knowledge-based methods obtained very good results with the MC30 and RG65 datasets, which can be explained by the deliberate inclusion of familiar and frequently used dictionary words in these sets.

Experiments - Result

- DataSet: Rubenstein and Goodenough(RG65), WordSimilarity-353(WS353), Mturk-771(MT771), Miller-Charles(MC30)

Pearson and Spearman results for the word relatedness datasets

Method		Pearson(γ)				Spearman(ρ)			
		MC30	RG65	WS353	MT771	MC30	RG65	WS353	MT771
Know-ledge-based	Wup	0.778	0.784	0.282	0.477	0.750	0.755	0.339	0.398
	J&C	0.695	0.731	0.354	0.498	0.820	0.804	0.318	0.402
	L&C	0.779	0.839	0.313	0.503	0.768	0.797	0.302	0.410
	Lin	0.835	0.858	0.329	0.513	0.750	0.788	0.348	0.424
	Resnik	0.813	0.836	0.362	0.431	0.693	0.731	0.353	0.404
Corpus-based	LSA	0.725	0.644	0.563	-	0.662	0.609	0.581	-
	ESA	0.588	-	0.503	-	0.727	-	0.629	-
	SSA	0.778	0.850	0.590	-	0.843	0.800	0.537	-
Ours	WTMGW	**0.879**	**0.861**	**0.622**	**0.572**	**0.846**	0.826	**0.750**	**0.480**
	WTMGR	0.871	0.847	0.602	0.539	0.820	**0.830**	0.748	0.477

As expected, our model performed better with large datasets such as WS353 and MT771, probably because the large datasets contained more technical and culturally biased terms, which cannot be covered by knowledge-based measures.

Experiments - Result

- DataSet: Lee50, Li30, AG400

Person and Spearman results for the text relatedness datasets							
Method		Pearson(γ)			Spearman(ρ)		
		Li30	Lee50	AG400	Li30	Lee50	AG400
Knowledge-based	COMB	0.810	0.702	0.480	0.832	0.356	0.365
Corpus-based	VSM	0.759	0.639	0.386	0.773	0.289	0.304
	LSA	0.810	0.635	0.425	0.812	0.437	0.389
	ESA	0.838	0.696	0.365	0.863	0.463	0.318
	SSA	0.848	0.684	0.567	0.832	0.480	0.495
Ours	WTMGW	**0.886**	**0.724**	**0.602**	**0.878**	**0.488**	0.486
	WTMGT	0.872	0.673	0.584	0.870	0.452	**0.512**

Our method clearly delivered the best performance, and they provided great improvements with the AG400 dataset.

But they were only slightly better than other methods with relatively small datasets

Conclusion

- We developed a word-text mutual guidance model to mine the deep relationships between words and texts, which combines semantic and statistical information using an iterative process.
- The experimental results demonstrated that our proposed model is more effective than the state-of-the-art methods for semantic computing.
- The evaluations using standard word-to-word and text-to-text relatedness benchmarks confirmed the superiority and consistency of our model.
- Drawback: the model remains time consuming.
- Improvement:
  - Possible to optimize the algorithm using dimensionality reduction.
  - Apply this semantic relatedness model to other NLP tasks such as text clustering or relationship classification.