Semantic Similarity

Marco Moresi
Dialog Systems and Machine Learning Group

26.06.2020

Content

- Semantic Similarity
  - What is semantic similarity?
  - Why do we need semantic similarity?
- Knowledge-based methods
- Corpus-based methods
- Deep Neural Network-based methods
- Transformer-based methods
- Conclusion

Semantic Similarity

- Semantic Textual Similarity (STS) is defined as the measure of semantic equivalence between two blocks of text.
- Semantic similarity methods usually give a ranking or percentage of similarity between texts, rather than a binary decision (similar or not).
- The versatility of natural language makes it difficult to define rule-based methods for determining semantic similarity.

Semantic Similarity

Heinrich Heine
Universität Düsseldorf

Where is semantic similarity used?

- Information retrieval
- Text summarization
- Text classification
- Essay evaluation
- Machine translation
- Question answering
- Natural language generation
- Spoken dialog systems

Semantic Similarity

Heinrich Heine
Universität Düsseldorf

First Approach

Semantic Similarity

First Approach

- Bag of Words (BoW)
  - Fixed vocabulary
  - Lose sequence order

Semantic Similarity

First Approach

- Bag of Words (BoW)

Example 1

Sentence 1: ”John and David studied Maths and Science.”

Sentence 2: “John studied Maths and David studied Science.”.

Semantic Similarity

Heinrich Heine Universität Düsseldorf

First Approach

- Bag of Words (BoW)

Example 1

Sentence 1: ”John and David studied Maths and Science.”

Sentence 2: “John studied Maths and David studied Science.”.

BoW Sentence1: {John: 1, David: 1, studied: 1, Maths: 1, Science: 1, and :2}
[1,1,1,1,1,2]

BoW Sentence2: {John: 1, David: 1, studied: 2, Maths: 1, Science: 1, and :1}
[1,1,2,1,1,1]

Semantic Similarity

Heinrich Heine Universität Düsseldorf

First Approach

- Bag of Words (BoW)

Example 1

Sentence 1: “John and David studied Maths and Science.”

Sentence 2: “John studied Maths and David studied Science.”.

Example 2

Sentence 1: “Mary is allergic to dairy products.”

BOW Sentence 1: {Mary: 1, is: 1, allergic: 1, to: 1, dairy: 1, products: 1, lactose: 0, intolerant: 0}
[1,1,1,1,1,1,0,0]

Sentence 2: “Mary is lactose intolerant.”

BOW Sentence 2: {Mary: 1, is: 1, allergic: 0, to: 0, dairy: 0, products: 0, lactose: 1, intolerant: 1}
[1,1,0,0,0,0,1,1]

Semantic Similarity

Heinrich Heine Universität Düsseldorf

First Approach

- Bag of Words (BoW)
  - Fixed vocabulary
  - Lose sequence order
- Term Frequency – Inverse document Frequency (TF-IDF)
  - TF measures how frequently a term occurs in a document.
  - IDF measures how important a term is.

Semantic Similarity

Heinrich Heine Universität Düsseldorf

First Approach

- Word Overlap
  - Calculated as a number of words that occur in both texts
- BLEU [Papineni et al., 2002]
  - Compare n-grams of the candidate with the n-grams of the reference
- ROUGE-L [Lin and Och, 2004]
  - Identifies longest co-occurring in sequence n-grams

Semantic Similarity

Knowledge-based methods

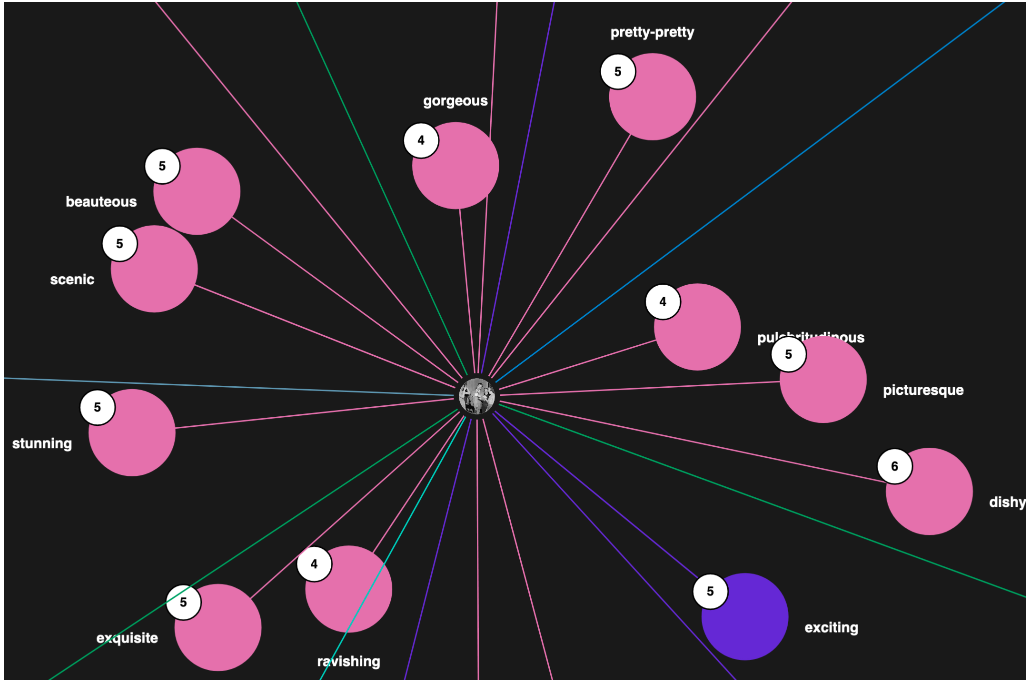
- Calculate semantic similarity between two terms based on the information derived from one or more underlying knowledge sources like ontologies/lexical databases, thesauri, dictionaries, etc
  - WordNet
  - Wiktionary
  - Wikipedia
  - BabelNet

Semantic Similarity

Heinrich Heine
Universität Düsseldorf

Knowledge-based methods

- BabelNet: It is the largest multilingual semantic ontology available with nearly over 13 million synsets and 380 million semantic relations.
- Synset: is a group of data elements that are considered semantically equivalent.



Synset of Beautiful (adj) in BabelNet
http://live.babelnet.org/

Semantic Similarity

Knowledge-based methods

- Edge-counting methods
- Feature-based methods
- Information Content-based methods

Semantic Similarity

Knowledge-based methods

- Edge-counting methods
  - Consider the underlying ontology as a graph, connecting words taxonomically.
  - The greater the distance between two terms the less similar they are.
- Feature-based methods
- Information Content-based methods

Semantic Similarity

Knowledge-based methods

- Edge-counting methods
  - Consider the underlying ontology as a graph, connecting words taxonomically.
  - The greater the distance between two terms the less similar they are.
- Feature-based methods
  - Calculate similarity as a function of properties of the words, like gloss.
    - Gloss, the meaning of a word in a dictionary.
  - Gloss-based semantic similarity
- Information Content-based methods

Semantic Similarity

Knowledge-based methods

- Edge-counting methods
  - Consider the underlying ontology as a graph, connecting words taxonomically.
  - The greater the distance between two terms the less similar they are.
- Feature-based methods
  - Calculate similarity as a function of properties of the words, like gloss.
  - Gloss-based semantic similarity
- Information Content-based methods
  - Information Content (IC)
  - Use the IC associated with the concept to evaluate similarity

Information Content

$$IC(c) = -\log p(c)$$

$$p(c) = \frac{\sum_{w \in W(c)} \text{appearances}\,(w)}{N}$$

$$\text{sim}_{res}(c_1, c_2) = IC(LCS(c_1, c_2))$$

Semantic Similarity

Corpus-based methods

- Word Embeddings

Semantic Similarity

Heinrich Heine
Universität Düsseldorf

Corpus-based methods

- Word Embeddings
  - word2vec
    - Neural network model
    - The Continuous Bag Of Words (CBOW) model predicts the current word using the previous words
    - The Skip-gram model predicts the neighboring context words given a target word.
  - GloVe
    - Word co-occurrence matrix
  - fastText
    - Skip-gram model
    - Each word is represented as a collection of character n-grams

Semantic Similarity

Corpus-based methods

- Word Embeddings
  - word2vec
  - GloVe
  - fastText

Meaning Conflation Problem

Bat

X = [0.50451 , 0.68607, …, -0.51042]

Semantic Similarity

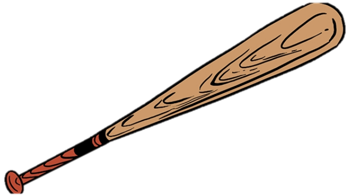
Corpus-based methods

- Word Embeddings
  - word2vec
  - GloVe
  - fastText

Meaning Conflation Problem

Bat

X = [0.50451 , 0.68607, …, -0.51042]



Semantic Similarity

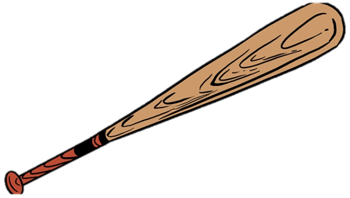
Corpus-based methods

- Word Embeddings
  - word2vec
  - GloVe
  - fastText

Meaning Conflation Problem

Bat

X = [0.50451 , 0.68607, …, -0.51042]



Semantic Similarity

Corpus-based methods

- Word Embeddings
  - word2vec
  - GloVe
  - fastText

Cosine Similarity

$$\text{similarity} = \cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\|\|\mathbf{B}\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}}$$

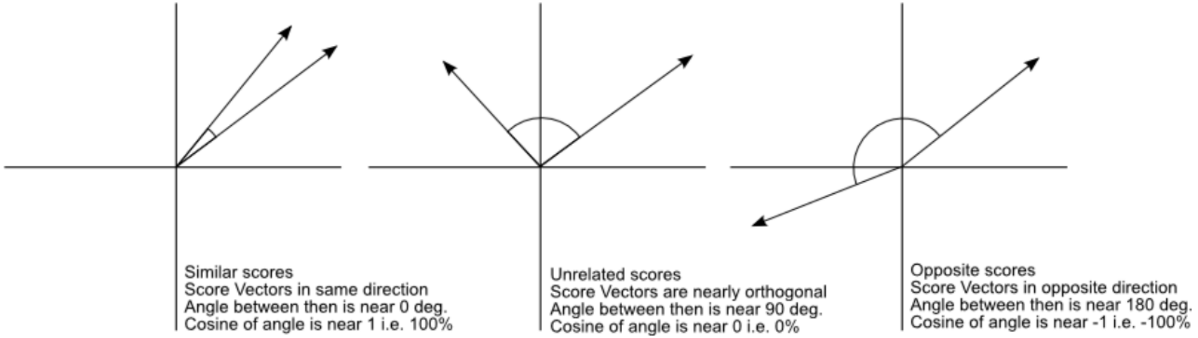
- Latent Semantic Analysis
  - Co-occurrence matrix, rows represent words and columns paragraphs
  - Singular Value Decomposition (SVD)
  - Each word is represented as a vector using the values in its row
  - Semantic Similarity is calculated using cosine similarity between these vectors

Semantic Similarity

Corpus-based methods

- Word Embeddings
  - word2vec
  - GloVe
  - fastText

- Latent Semantic Anal
  - Co-occurrence matrix
    columns paragraphs
  - Singular Value Decor
  - Each word is represented as a vector using the values in its row
  - Semantic Similarity is calculated using cosine similarity between these vectors

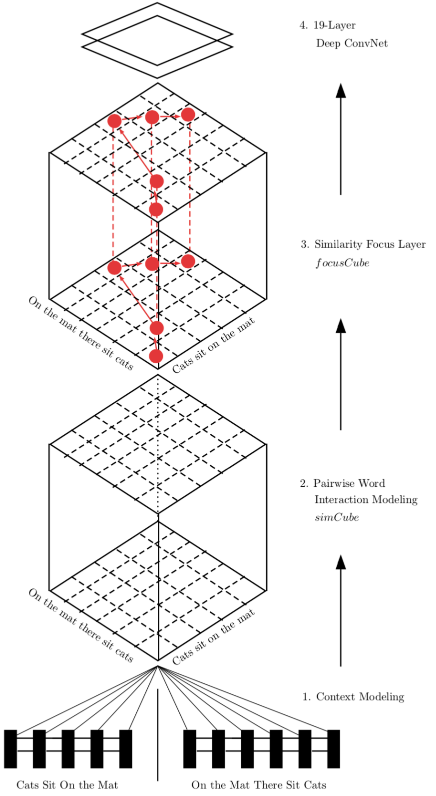


The Cosine Similarity values for different documents, 1 (same direction), 0 (90 deg.), -1 (opposite directions).

Semantic Similarity

Deep Neural Network-based methods

- Pairwise Word Interaction Modeling with Deep Neural Networks for Semantic Similarity Measurement
  - Context Modeling
    - BiLSTM to model the context
  - Pairwise Word Interaction Modeling
    - Establish semantic correspondence
  - Similarity Focus Layer
    - FocusCube
  - Deep ConvNet
    - FocusCube as an "image"
    - Pattern Recognition problem



Pairwise Word Interaction Modeling with Deep Neural Networks for Semantic Similarity Measurement. [He.et al 2016]

How good is the proposed metric?

Heinrich Heine
Universität Düsseldorf

How can we evaluate how good is the metric?

- Correlation with human annotation
  - We need humans to rank pair of sentences according how similar they are
  - Calculate the correlation between the proposed metric and the human annotations
  - Pearson Correlation

Semantic Similarity

Heinrich Heine
Universität Düsseldorf

Deep Neural Network-based methods

- Pairwise Word Interaction Modeling with Deep Neural Networks for Semantic Similarity Measurement

STS2014	3rd	2nd	1st	This work
deft-forum	0.5305	0.4711	0.4828	**0.5684**
deft-news	**0.7813**	0.7628	0.7657	0.7079
headlines	**0.7837**	0.7597	0.7646	0.7551
image	**0.8343**	0.8013	0.8214	0.8221
OnWN	0.8502	0.8745	0.8589	**0.8847**
tweetnews	0.6755	**0.7793**	0.7639	0.7469
Wt. Mean	0.7549	0.7605	0.761	**0.7666**

Test results on all six test sets in STS2014. Pearson's r scores calculated based on the number of sentence pairs in each test set

Pairwise Word Interaction Modeling with Deep Neural Networks for Semantic Similarity Measurement. [He.et al 2016]

Semantic Similarity

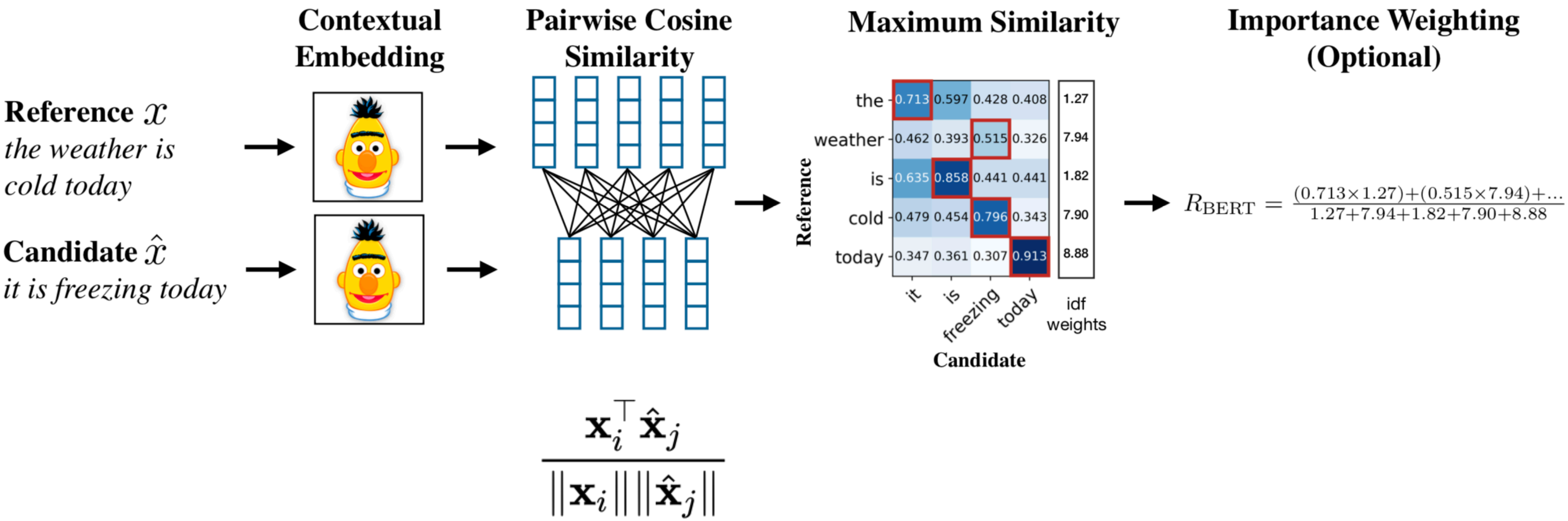
Transformer-based methods

- BERTScore: Evaluating text generation with Bert [Zhang et al. 2020]

Semantic Similarity

Transformer-based methods

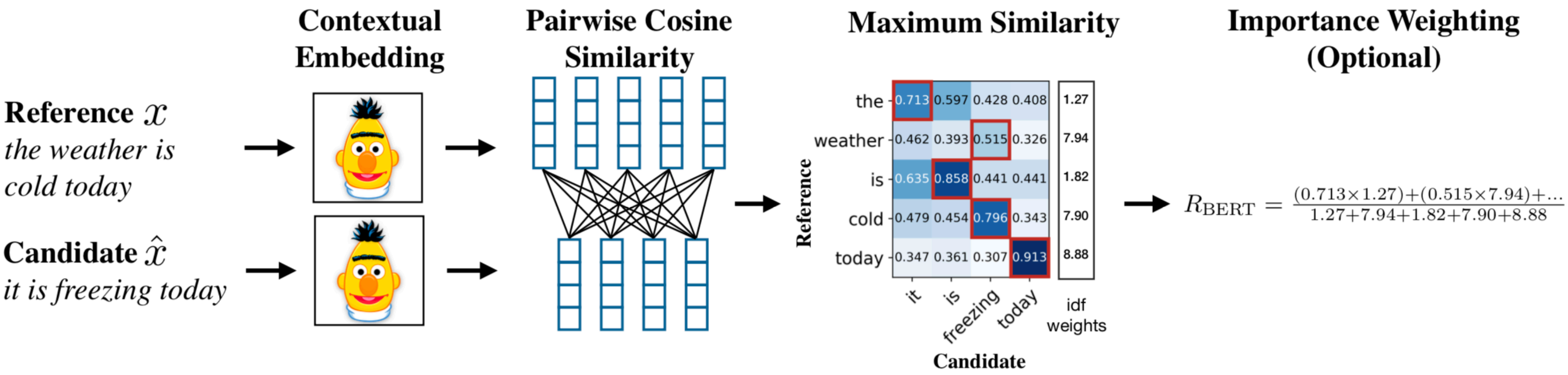
- BertScore



Semantic Similarity

Transformer-based methods

- BertScore

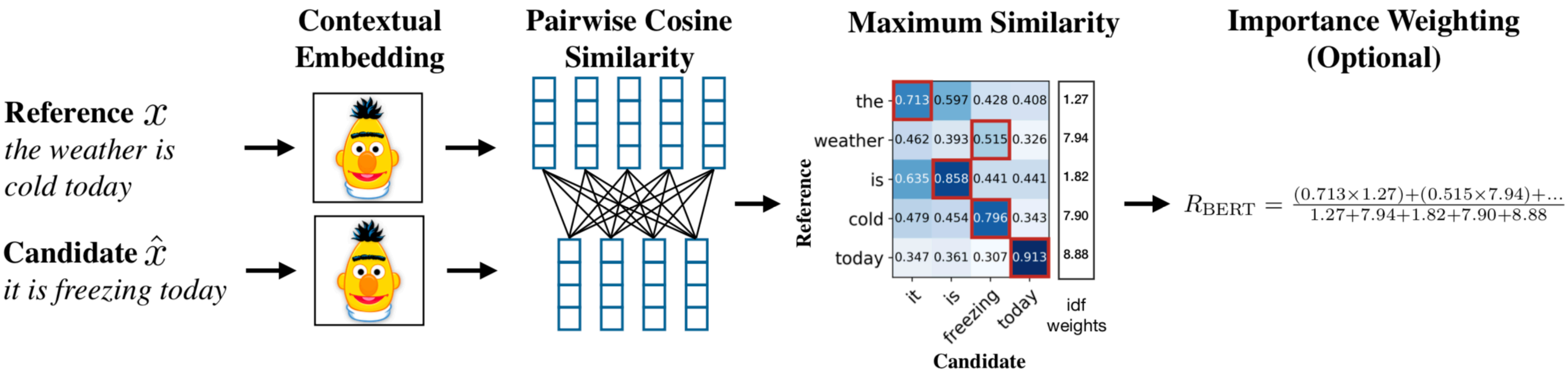


$$\mathrm{idf}(w) = -\log \frac{1}{M} \sum_{i=1}^{M} \mathbb{I}\left[w \in x^{(i)}\right]$$

Semantic Similarity

Transformer-based methods

- BertScore

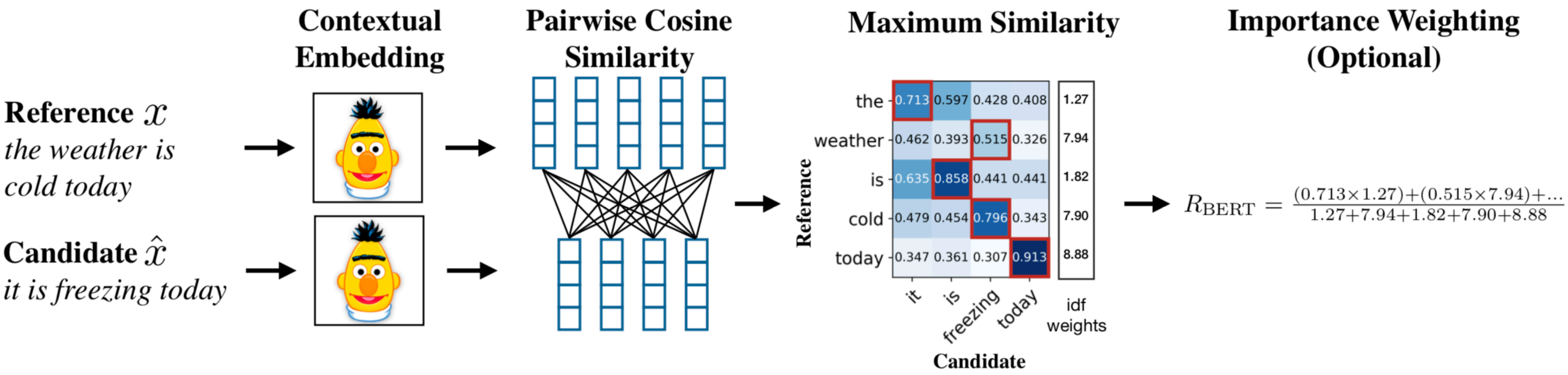


$$R_{\mathrm{BERT}} = \frac{\sum_{x_i \in x} \mathrm{idf}(x_i) \max_{\hat{x}_j \in \hat{x}} \mathbf{x}_i^\top \hat{\mathbf{x}}_j}{\sum_{x_i \in x} \mathrm{idf}(x_i)}$$

Semantic Similarity

Transformer-based methods

- BertScore



$$F_{\mathrm{BERT}} = 2\frac{P_{\mathrm{BERT}} \cdot R_{\mathrm{BERT}}}{P_{\mathrm{BERT}} + R_{\mathrm{BERT}}}$$

Semantic Similarity

Heinrich Heine
Universität Düsseldorf

Transformer-based methods

- BertScore

METRIC	en↔cs	en↔de	en↔fi	en↔zh
BLEU	.956/.983	.969/.977	.962/.958	.968/.941
P_{BERT}	.965/.989	.995/.983	.976/.951	.975/.950
R_{BERT}	**.989/.995**	**.997/.991**	**.989/.977**	**.981/.980**
F_{BERT}	.978/.993	.989/.978	.984/.969	.981/.969
$F_{BERT(idf)}$	.982/.995	.988/.979	.989/.969	.980/.963

Pearson correlation. WMT18 dataset, translation pairs, English(en) to Chinese(cs), German(de), Finish(fi) and Czech(zh).
the left number is the to-English correlation, and the right is the from-English.

Conclusion

hhu
Heinrich Heine
Universität Düsseldorf

- Measuring semantic similarity between two snippets of text is one of the most challenging tasks in Natural Language Processing.

- Knowledge-based models: consider the meaning of the text but are not adaptable across different domains and languages.

- Corpus-based models: have a statistical background and can be implemented across languages Do not consider the meaning of the text.

- Deep Neural Network based models: show better performance but require high computational resources.

- Transformer based models: take advantage of the pre-training, contextual embedding, are of the state of the art.

Heinrich Heine
Universität Düsseldorf

Questions?

Heinrich Heine
Universität Düsseldorf

Thank you!

References

- BERTScore: Evaluating text generation with Bert [Zhang et al. 2020]
- Pairwise Word Interaction Modeling with Deep Neural Networks for Semantic Similarity Measurement [He. et al 2014]
- ROUGE: A Package for Automatic Evaluation of Summaries [Lin, 2004]
- BLEU: a Method for Automatic Evaluation of MachineTranslation [Papineni et al. 2002]
- Introduction to WordNet: An On-line Lexical Database [Miller et al. 1993]
- BabelNet: The automatic construction, evaluation and applicationof a wide-coverage multilingual semantic network [Navigli et al 2012]