Sentence Similarity Measures for Fine-Grained Estimation of Topical Relevance in Learner Essays

Marek Rei and Ronan Cummins
ALTA Institute
Computer Laboratory
UNIVERSITY OF CAMBRIDGE

Detecting the topical relevance of learner essays

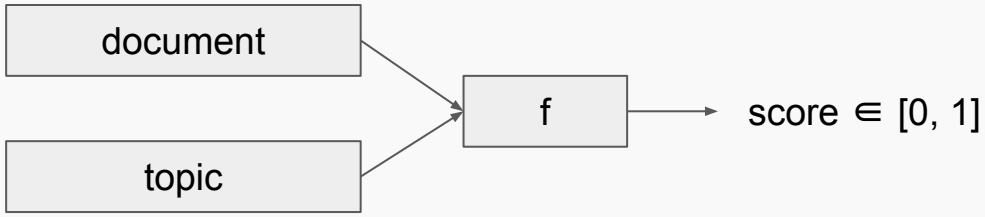
Motivation for topic relevance detection:

- Detect unsuitable topic shifts
- Detect memorised responses

Can train a topic-specific classifier to detect relevant texts.

but we need a training set for each topic.

Can construct a topic-independent scoring function to detect relevance between the topic and the text.



can use it on previously unseen topics.

Sentence-level topic relevance

- Able to provide more **fine-grained** feedback.

When discussing the topic of prison system , it is clear that the subject is vast and diversified if all different systems and circumstances that exist are taken into consideration . So I intend to focus on the situation mainly from the Finnish point of view. The prison system has developed over the past decades , let alone the past centuries . In modern prisons circumstances are more human : cells (or rooms) are normally fairly comfortable , often having conveniences such as television , coffee-machine , personal things etc. Prisoners have better possibilities for hobbies , studying and meeting visitors . No longer are prisoners kept on bread and water , though these kind of expressions stay in language . In general , people usually have a little knowledge of how prisons operate ; it is only a small minority of people who ever come into contact with the prison system. There are several disadvantages in the prison system . Firstly , the time spent in an oppressive and tough atmosphere of a prison often has a bad influence on prisoners , especially on young ones .

- Can be used for estimating the **coherence** of an essay.

- Can be used as a feature for **sentence quality** estimation (Andersen et al., 2013).

TF-IDF (Sparck Jones, 1972)

We can map sentences and prompts to vectors and measure their cosine similarity.

TF-IDF over words to construct vector representations for the topic and the target sentence.

$$IDF(w) = log(\frac{N}{1 + n_w})$$

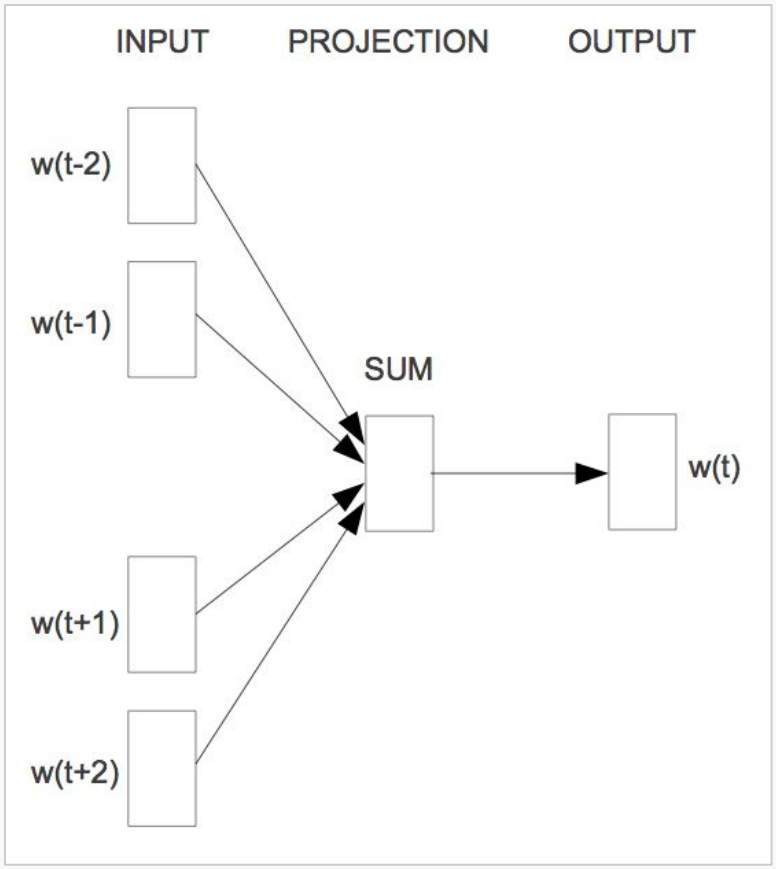
Assigns low weights to frequent words (determiners, prepositions, etc).

Assigns high weights to rare words (often spcecific content words).

Word frequency statistics collected from 100M words in the BNC.

Word2vec (CBOW, Mikolov et al, 2015)

- Learns distributed vector representations.
- Trains the vectors of the context words to predict the target word.
- To create a sentence vector, we add together the vectors for all the words in that sentence.
- We use the publicly available vectors, trained on 100B words of news text.



IDF-Embeddings

Hypothesis: we can improve this additive model by individually weighting each word.

Let's scale each word embedding with the IDF weight of the corresponding word.

$$\vec{w}' = IDF(w) \cdot \vec{w}$$

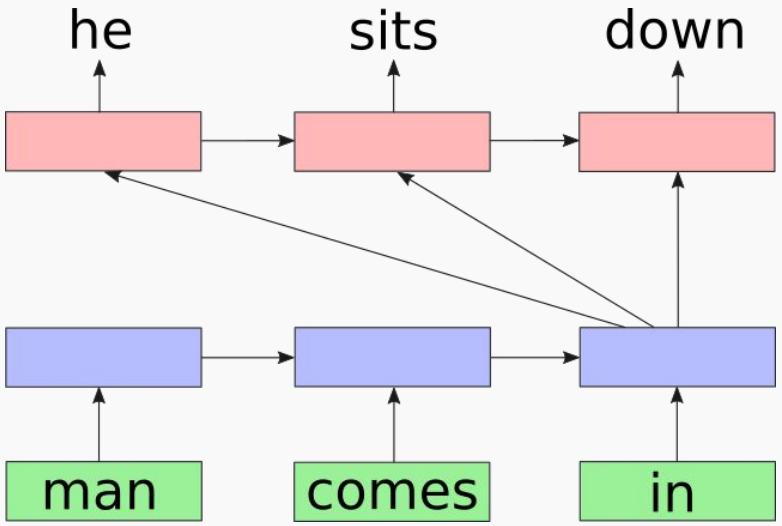
Retains the direction of each embedding.

But more frequent words now have lower impact on the sum.

Skip-Thoughts (Kiros et al., 2015)

A sentence is mapped to a vector using a recurrent network.

The model is trained to predict words in the surrounding sentences, conditioned on that sentence vector.



Trained on 985M words from unpublished books.

Weighted-Embeddings

Scale word embeddings with a weight, which we learn automatically from data.

1. Pick a main sentence **u**
2. Pick a nearby sentence **v** (which is likely to be related to u)
3. Pick a random sentence **z**
4. Construct sentence vectors by summing weighted word embeddings
5. Optimise the word weights $\mathbf{g_w}$ so that u and v are similar, and u and z are dissimilar.

$$\vec{u} = \sum_{w \in u} g_w \vec{w}$$

$$cost = max(-\vec{u}\vec{v} + \vec{u}\vec{z}, 0)$$

Evaluation

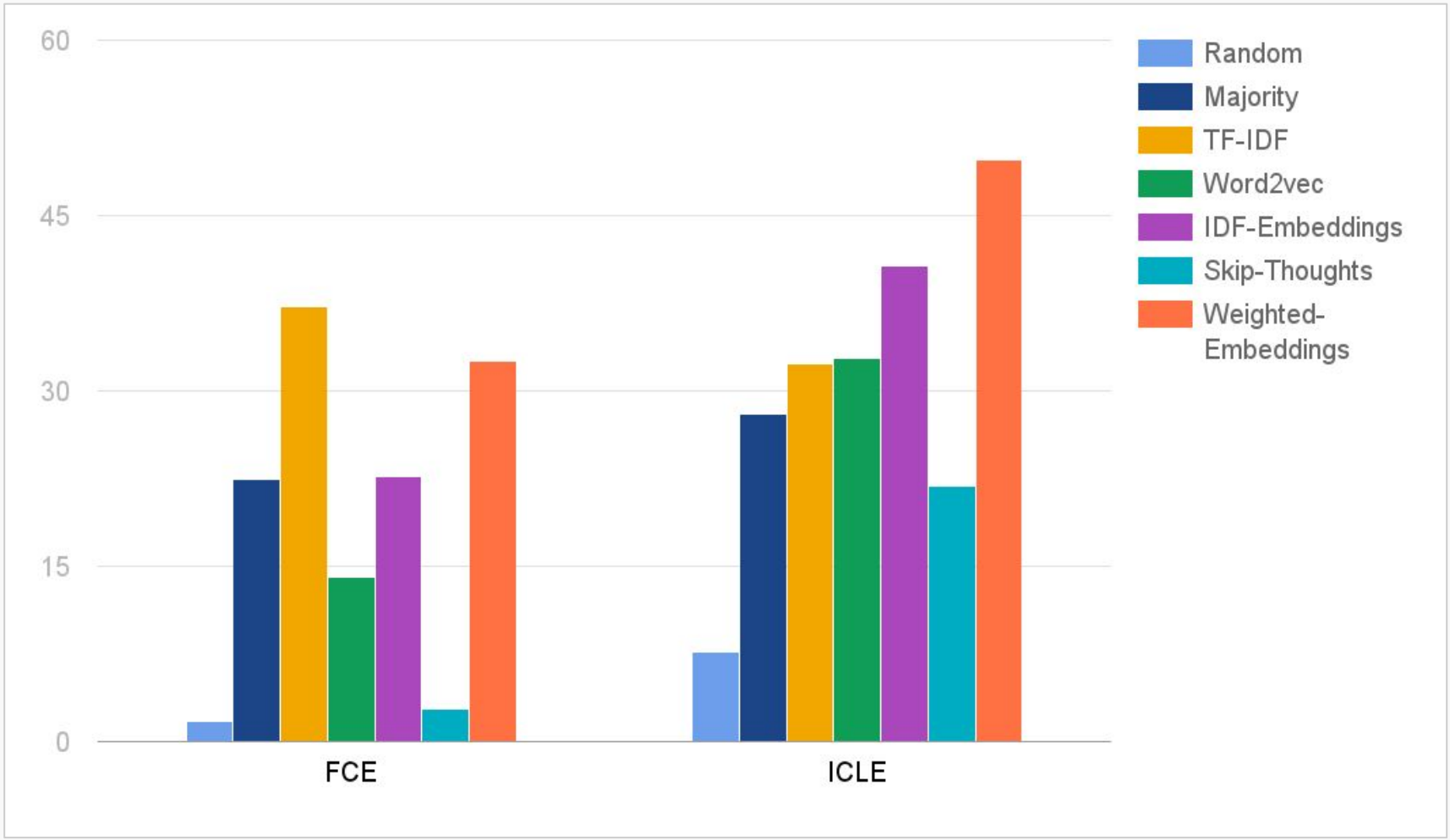
Using two publicly available corpora of learner essays:

1. First Certificate in English (FCE, Yannakoudakis et al. 2011)
   30,899 sentences and 60 prompts
   Detailed prompts, describing a scenario or giving instructions on what to mention in the text. Average prompt has 10.3 sentences.

2. International Corpus of Learner English (ICLE, Granger et al. 2009)
   20,883 sentences and 13 prompts.
   Short and general prompts, designed to point the student towards an open discussion around a topic. Average prompt has 1.5 sentences.

The system is presented with each sentence independently and it aims to correctly identify the prompt that the student was following.

Results: accuracy



Example output

Most University degrees are theoretical and do not prepare us for the real life. Do you agree or disagree?

0.382	Students have to study subjects which are not closely related to the subject they want to specialize in.
0.329	In order for that to happen however, our government has to offer more and more jobs for students.
0.085	I thought the time had stopped and the day on which the results had to be announced never came.

Most relevant words for this prompt:

University, degrees, undergraduate, doctorate, professors, university, degree, professor, PhD, College, psychology

Example weights

cos	3.32	two	-1.31
studio	2.22	although	-1.26
Labour	2.18	which	-1.09
want	2.01	five	-1.06
US	2.00	during	-0.80
Secretary	1.99	the	-0.73
Ref	1.98	unless	-0.66
film	1.98	since	-0.66
v.	1.91	when	-0.66
Cup	1.89	also	-0.65
data	1.88	being	-0.63
drink	1.88	high	-0.62
Minister	1.87	especially	-0.62
IBM	1.86	their	-0.62
Act	1.86	making	-0.61

Conclusion

- We can measure topic relevance of learner essays at the sentence level, using an unsupervised similarity function.
- TF-IDF is the best measure when the prompts are highly detailed.
- Embeddings-based methods are best when the prompts are short and general.
- We can improve embedding-based vectors by learning the individual weights for each word.
- By optimising the model for sentence similarity, the weights learn to assign higher importance to topic-specific words.

Thank you!