Automated Assessment of Text Comprehension Tasks

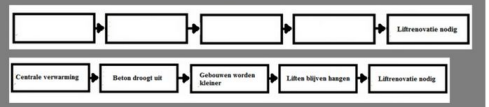
Karol Skalski
Supervisor: Dr Anique de Bruin

Plan of the presentation

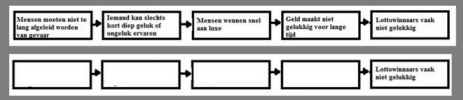
1. The problem
    a. The design of the original experiment
    b. Reformulation of the problem
    c. The Method used
2. A step back: What are word embeddings
    a. Assumptions of word embeddings
    b. Comparison methods
3. Application to the dataset
4. Results
5. Future possibilities
6. Tools used

DIAGRAM COMPLETION TASKS TO BE GRADED:

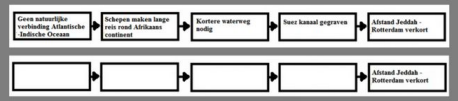
Renovatie van betonnen gebouwen



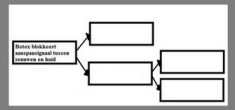
Geld maakt niet gelukkig



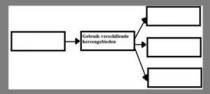
Het Suezkanaal



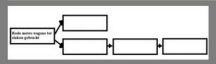
Gebruik van Botox



Muziek maakt slimmer



Metrowagons tot zinken gebracht

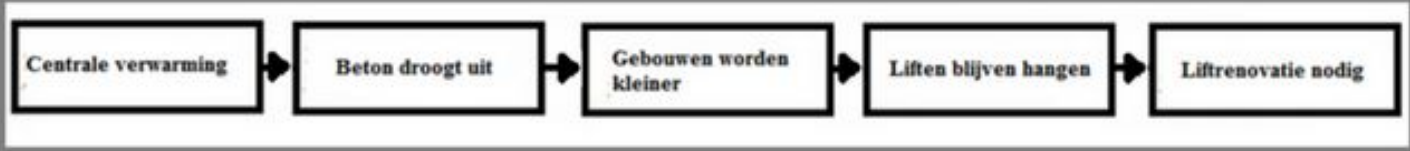


CODE CREDIT: NLP-TOWN@GITHUB

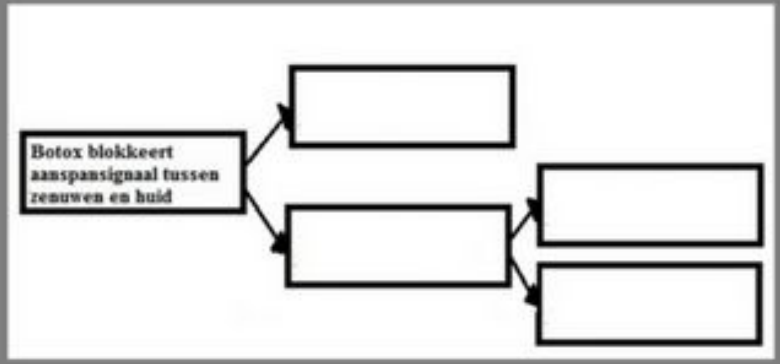
Code is not complete,
full script available on request
email me:
k.skalski@student.maastrichtuniversity.nl

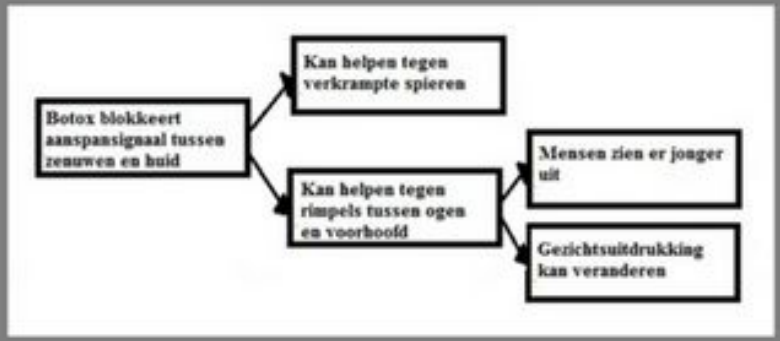
DIAGRAM COMPLETION TASKS TO BE GRADED:

Renovatie van betonnen gebouwen

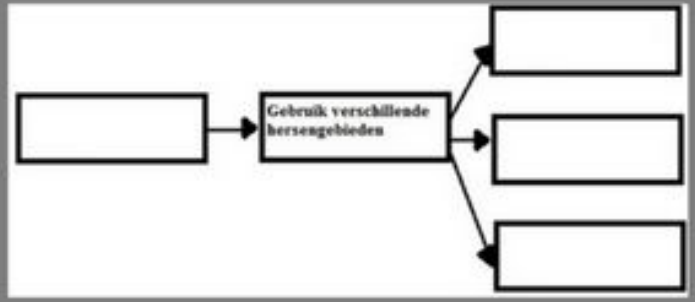


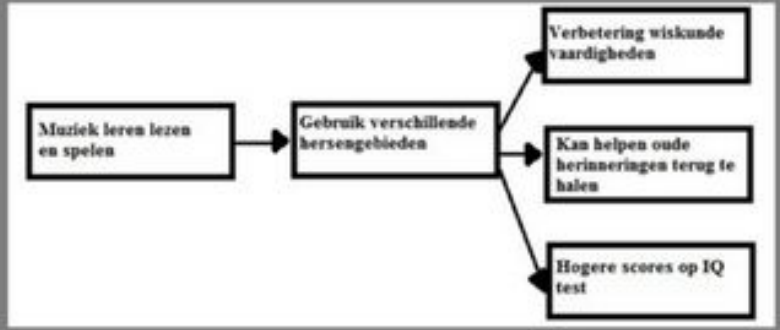
Gebruik van Botox





Muziek maakt slimmer





- **4 fields**
- **1 model answer per field**
- **Fields arranged in a sequence**

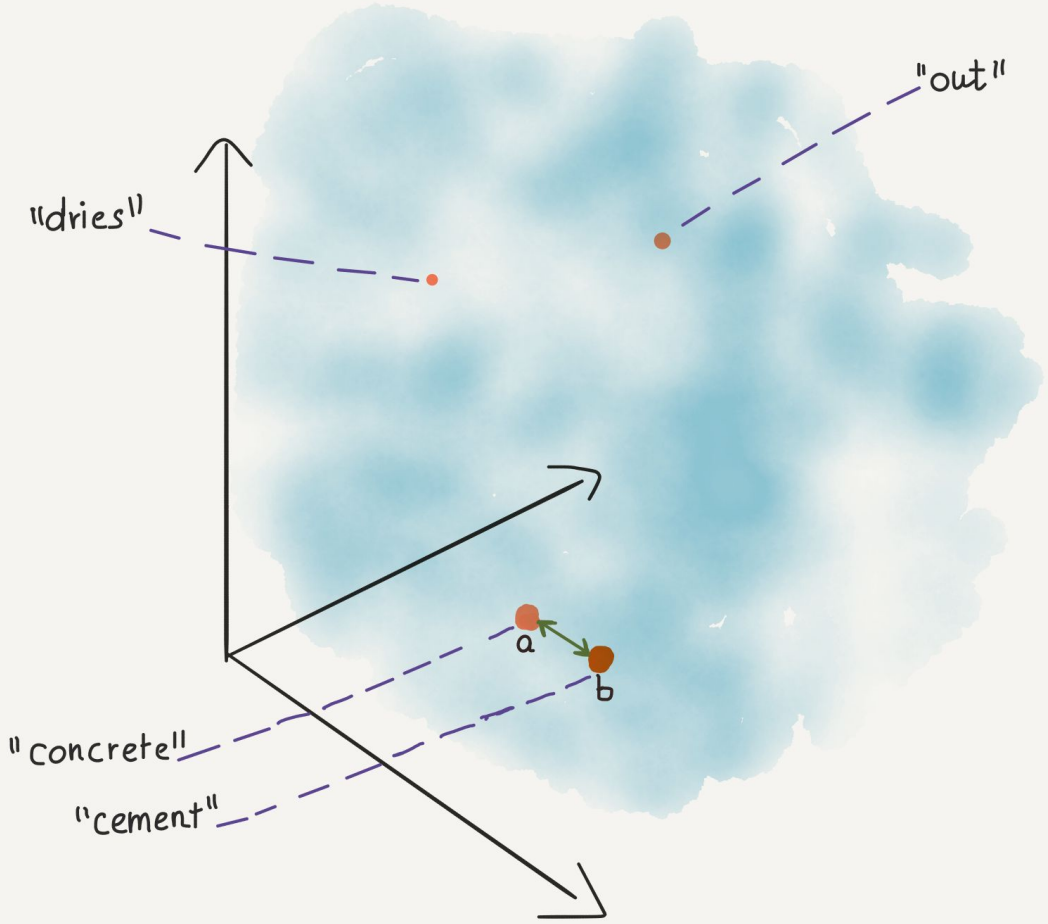
Already existing assessment metrics

- Accuracy score
- #correct_relations
- Self assessment

Default approach: For each task compare the response in field x to the model response in field x.

The problem reformulated into: How to measure (quantify) a semantic similarity between two sentences: original response and model response

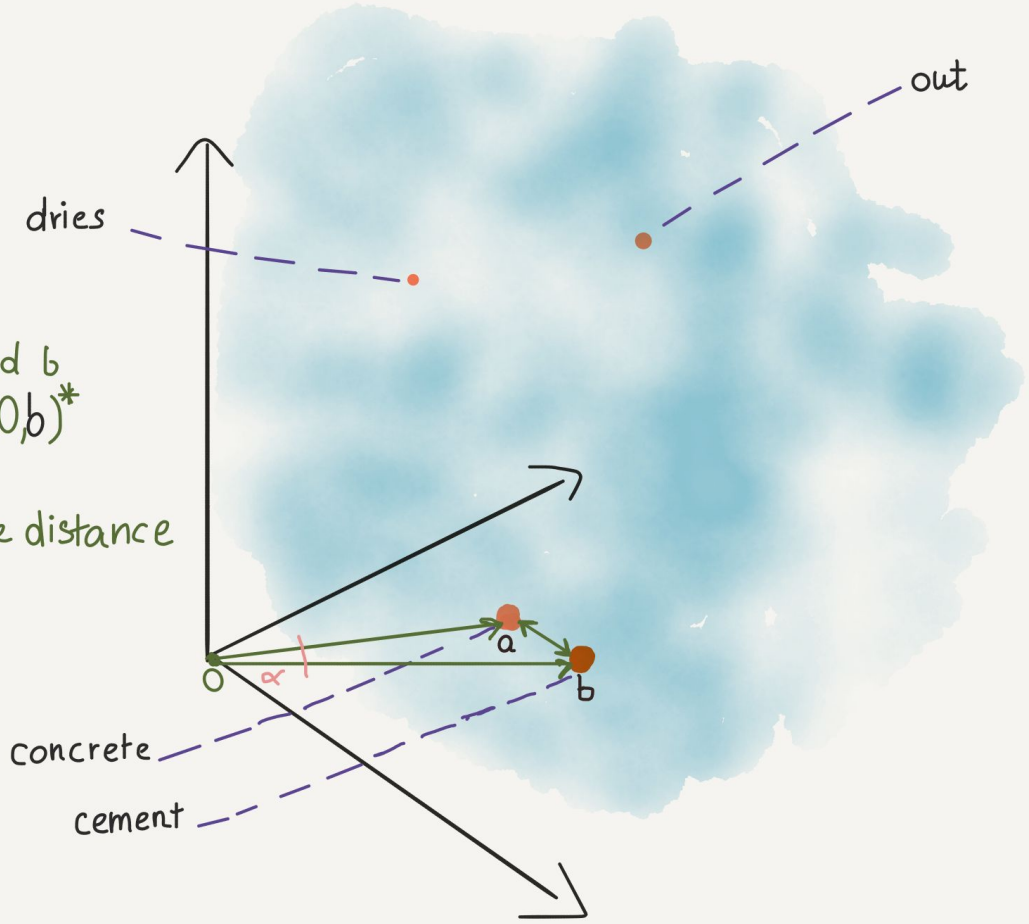
One level lower - how to quantify the semantic similarity between two words?


"out"
"dries"
a
b
"concrete"
"cement"

closeness of a and b

$c(a, b) = \cos(\angle a, 0, b)^*$

not the negative distance between words.



in the same text

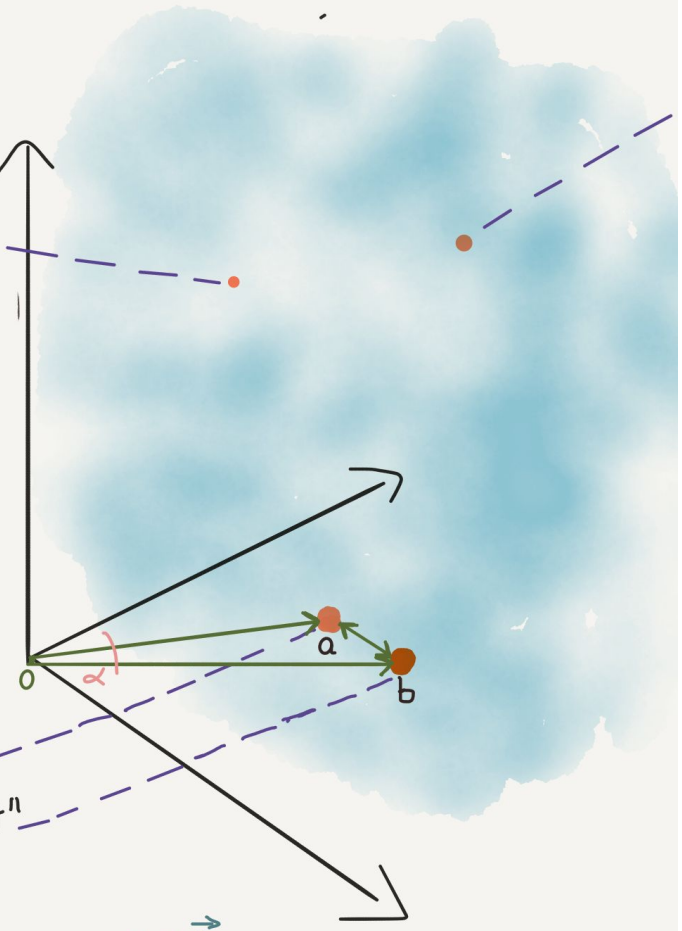
$P(a|b) = P(\boxed{\text{cement}} \text{ given } \boxed{\text{concrete}})$

$c(a,b) = \text{closeness}(a,b) = \cos(\measuredangle a,0,b)$

$Err_{a,b} = |P(a|b) - c(a,b)|^{*}$

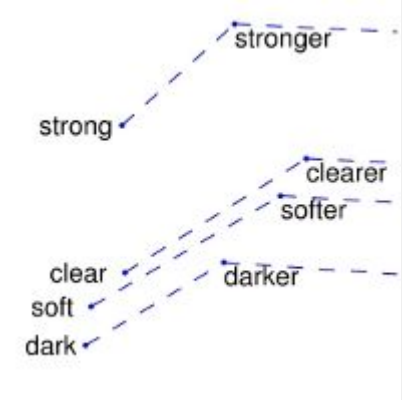
$$Err = \sum_{a,b} |1 - P(a|b) - c(a,b)|$$

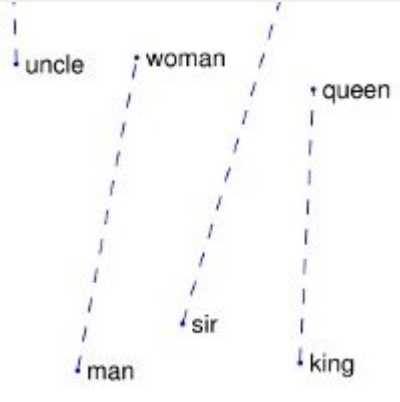
sum up for all "a"s and "b"s

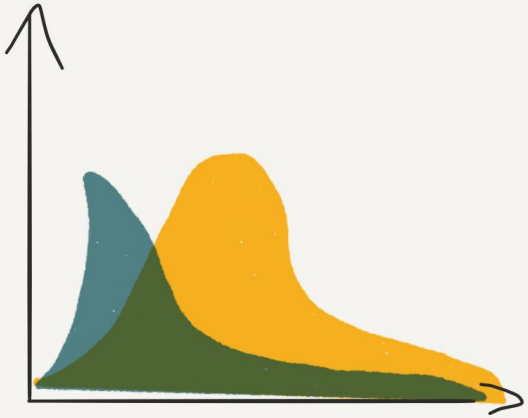


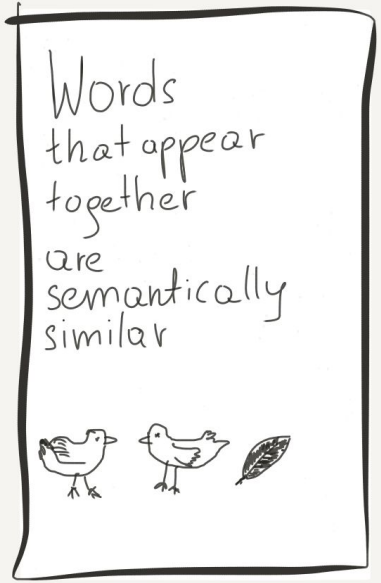
* in fact during trainng closeness is measured by a dot product of $\vec{a} \cdot \vec{b}$, than softmax classifier is used to obtain expected probabilities

Distributional hypothesis

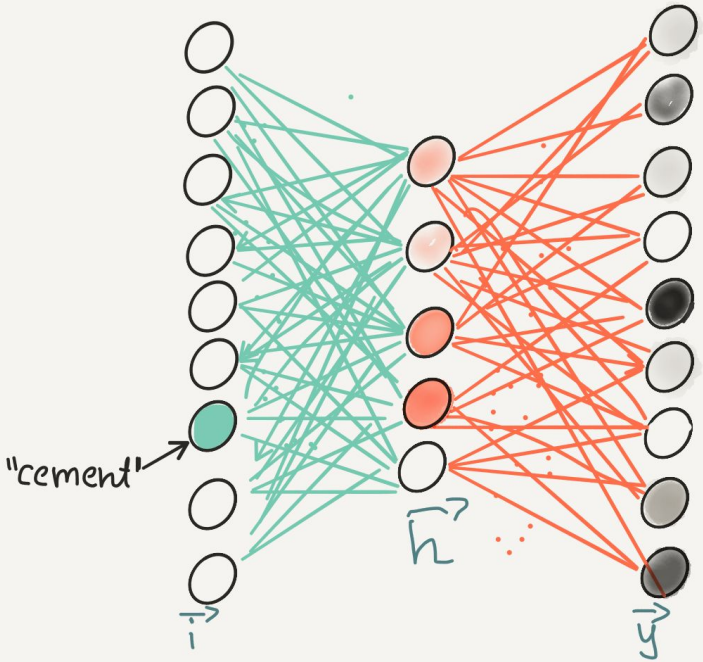




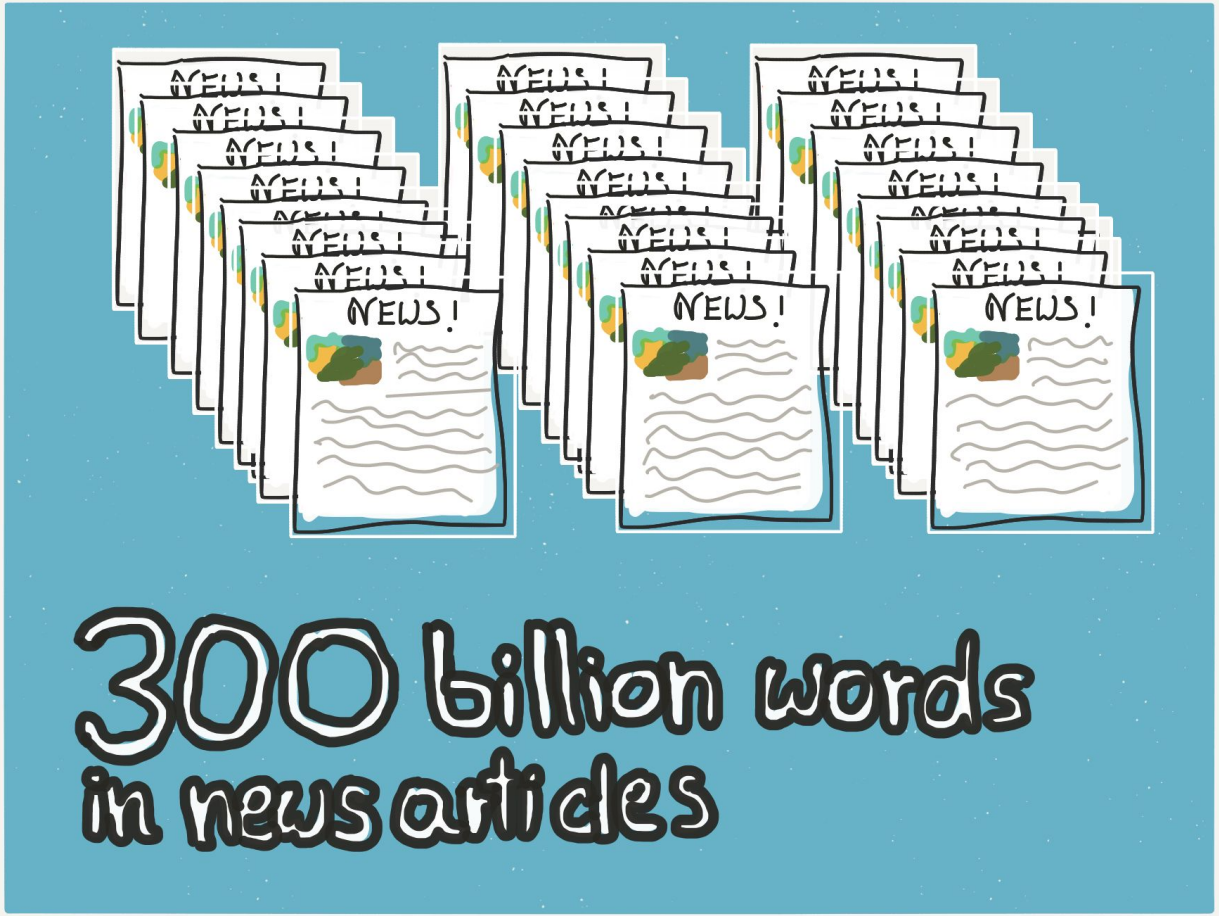


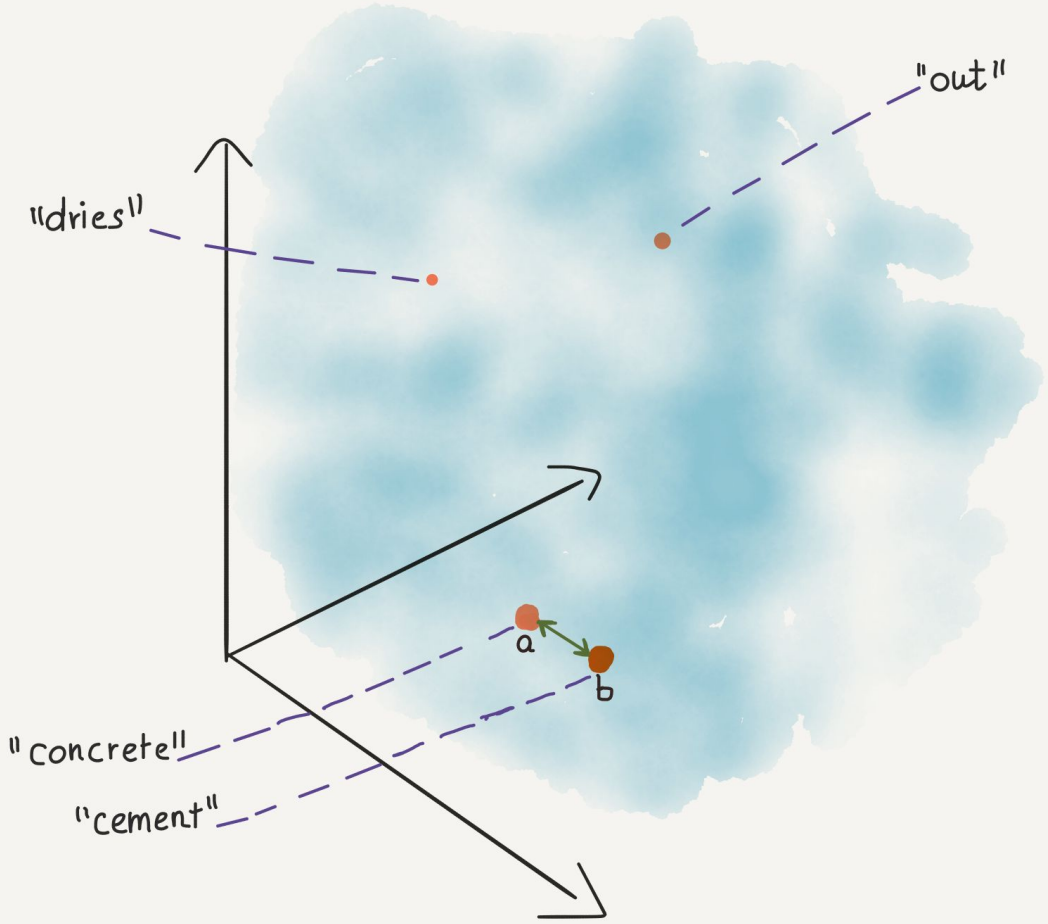


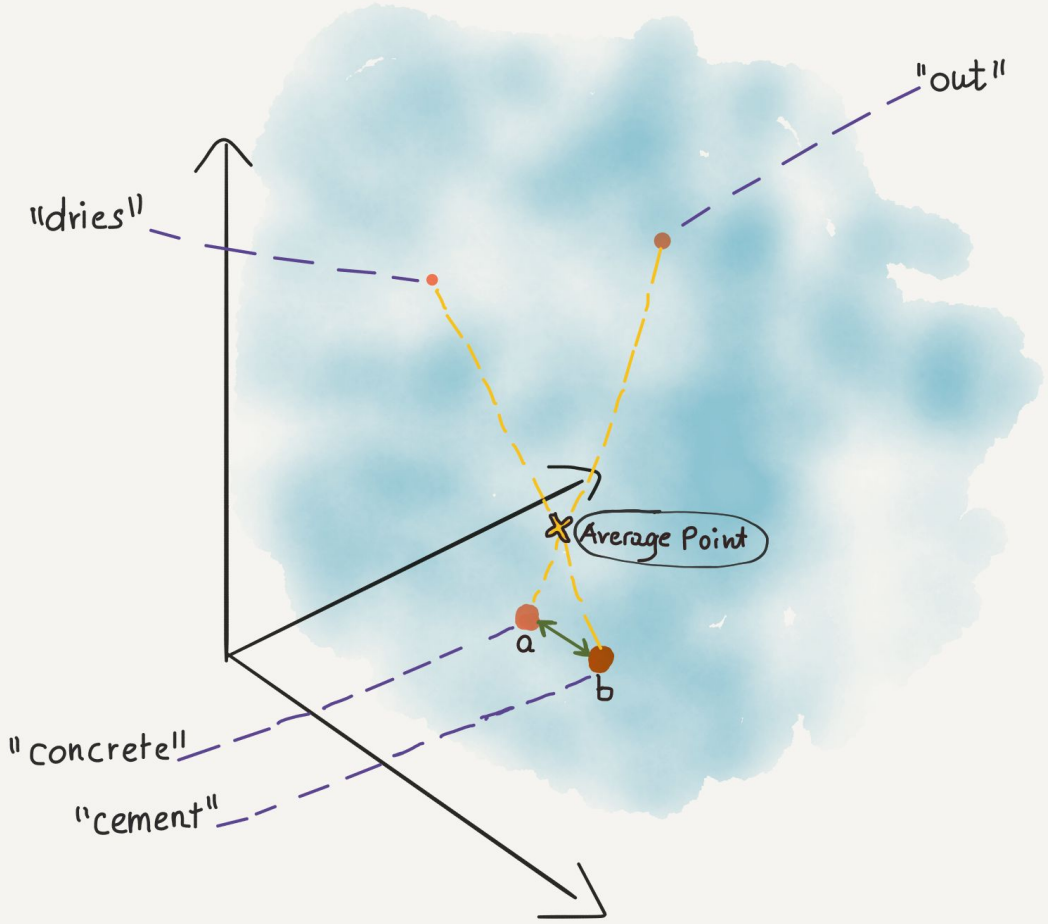
Word2vec



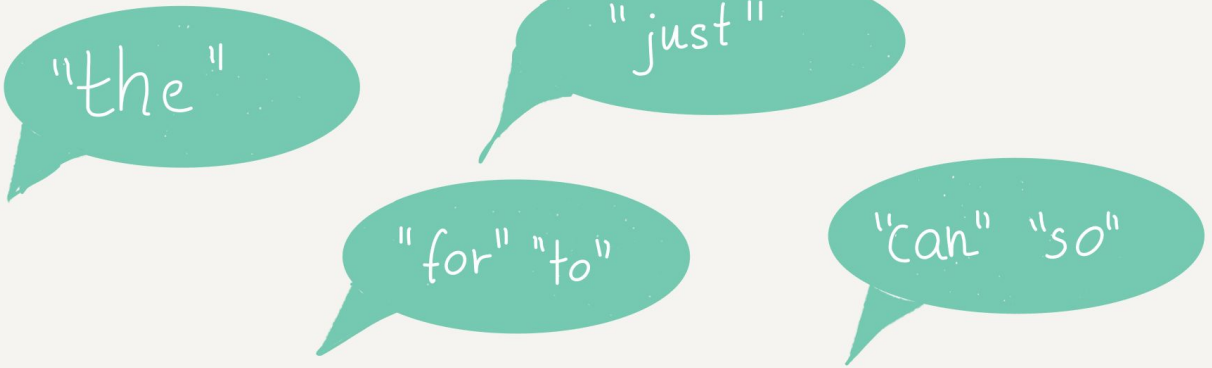
a vector representation of words trained with neural networks

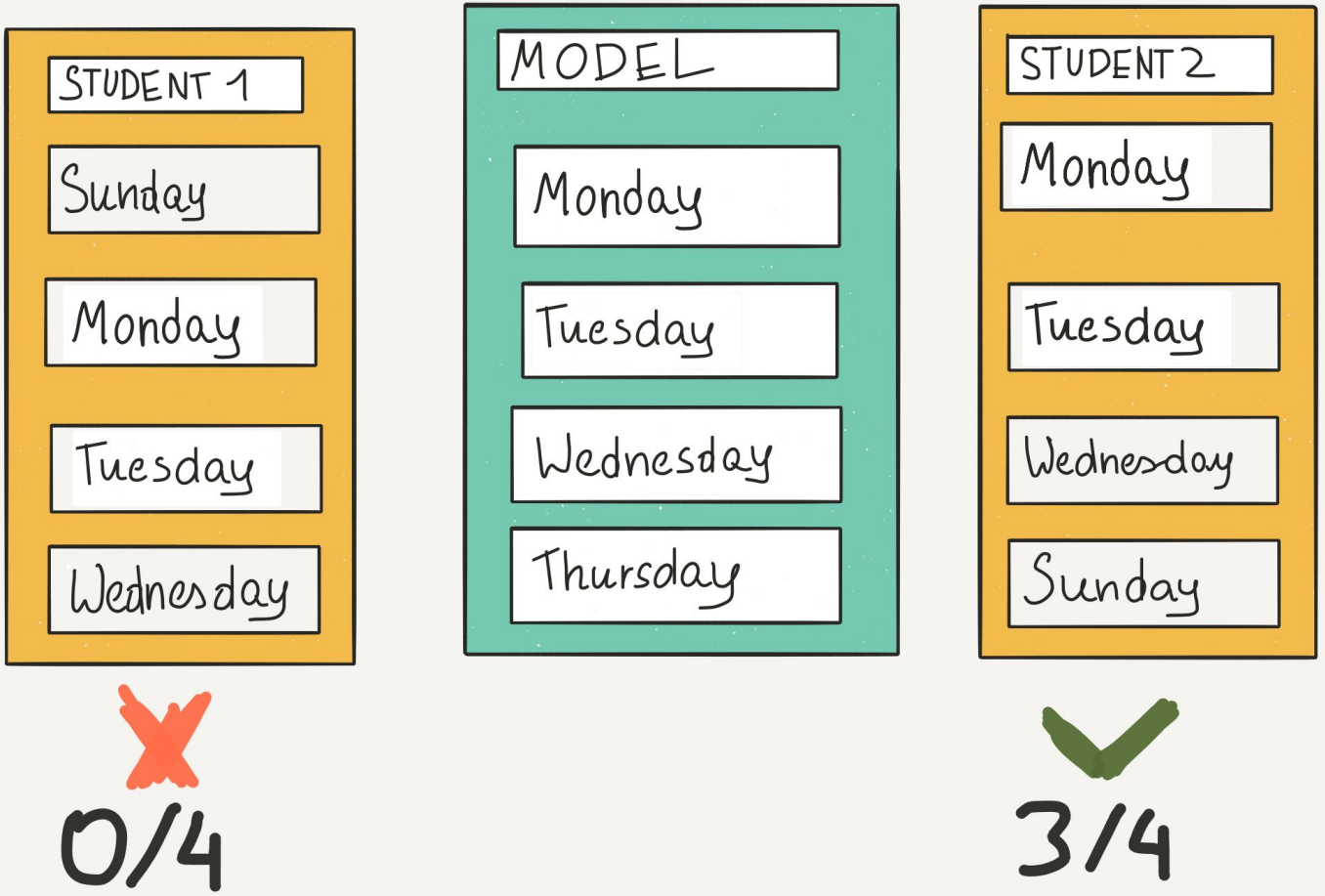

NEWS!
300 billion words
in news articles

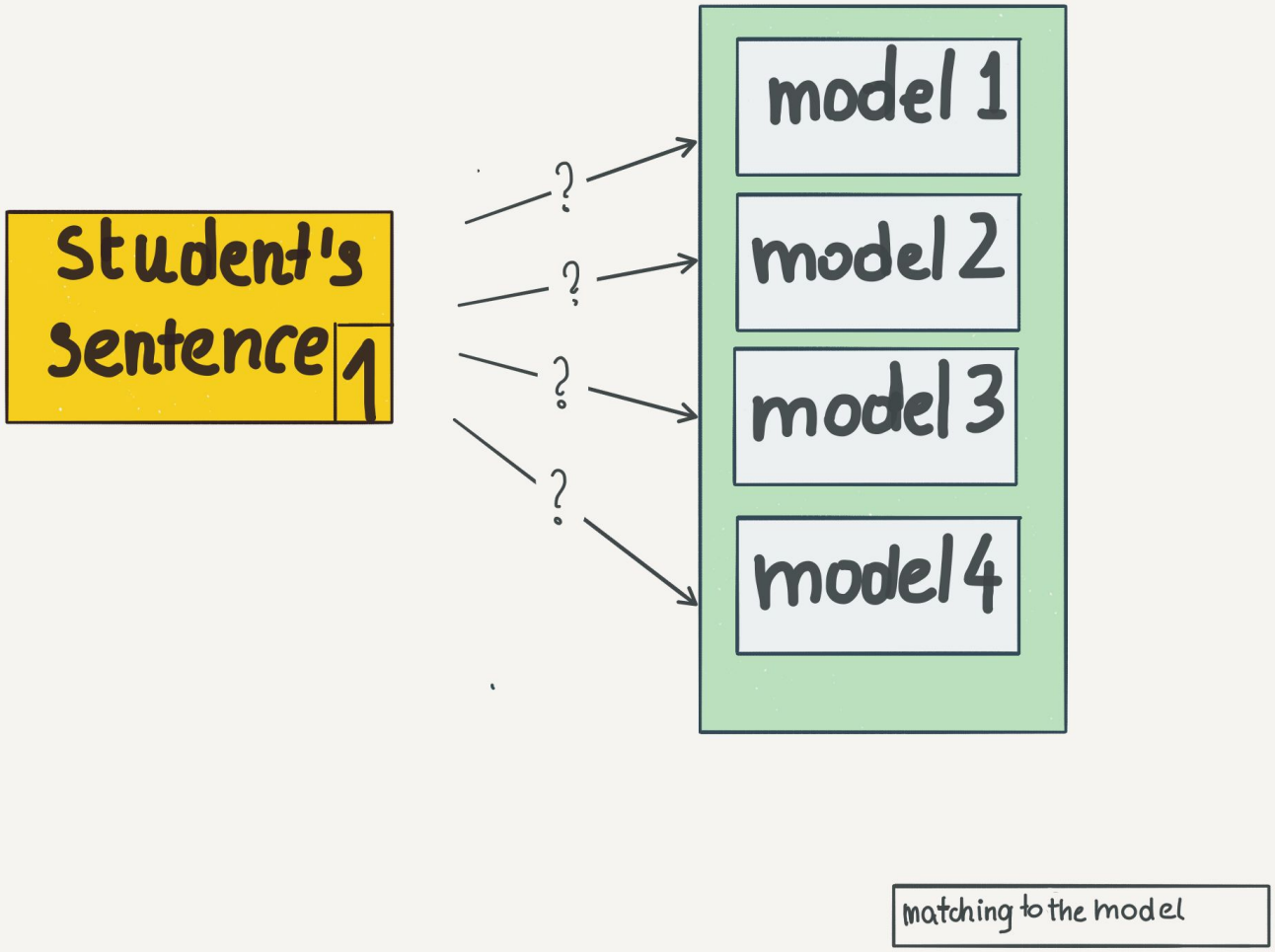

"out"
"dries"
a
b
"concrete"
"cement"


"out"
"dries"
Average Point
a
b
"concrete"
"cement"

Eliminate the STOP words




STUDENT 1
Sunday
Monday
Tuesday
Wednesday
0/4
MODEL
Monday
Tuesday
Wednesday
Thursday
STUDENT 2
Monday
Tuesday
Wednesday
Sunday
3/4

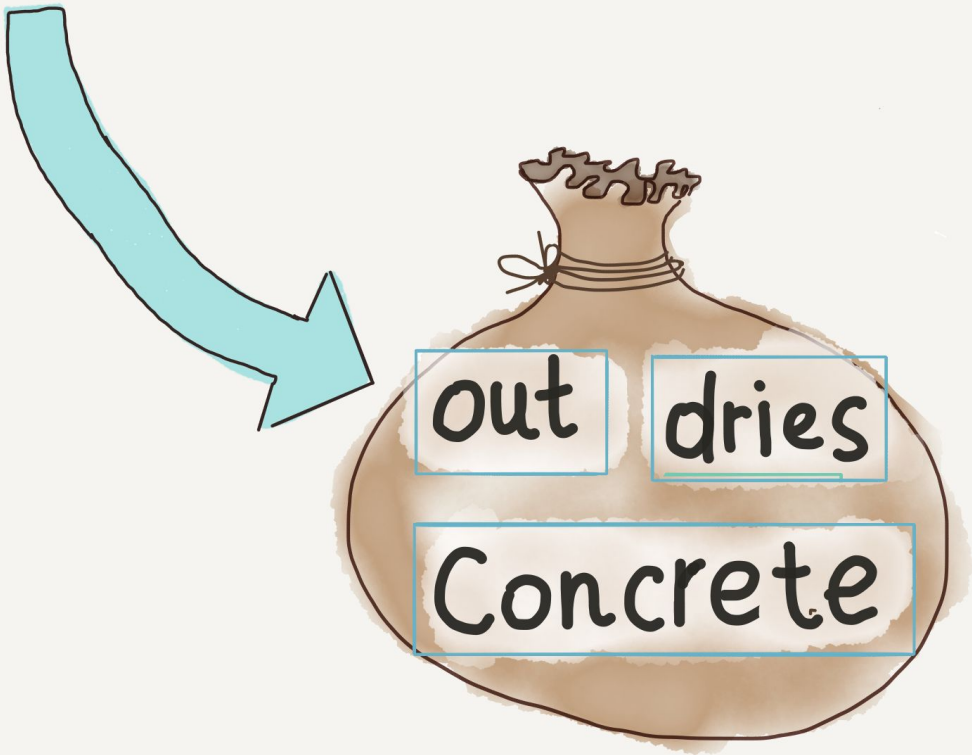

Student's Sentence 1
?
?
?
?
model 1
model 2
model 3
model 4
matching to the model

Levenshtein distance

"abcd"  "adbc"

"1324"  "1234"

Concrete dries out



The obvious issue:

(1) “The hare beats the tortoise.”

(2)“The tortoise beats the hare.”

1 and 2 are opposite in terms of meaning, but same in terms of context

Student's responses

4	2	By heating the concrete shrinks
5	3	concrete buildings shrink by central heatir
6	4	concrete decreases after a while
7	5	can shrink concrete
8	6	shrink concrete buildings
9	7	Buildings made of concrete but people dic
10	8	Concrete shrinks as it gets hot
11	9	v concrete buildings become smaller heat
12	10	get small concrete as it dries

Why bag of words solution is still preferred?

Model

Central heating
Can help prevent muscle tightness
People should not be distracted for too long risk
Waste need not be processed in some way
learn to read and play music
No natural connection Western Indian Ocean

The core assumption:

If students are talking about the right thing,

Means: the context of their response is the same as the model's.

They are saying the right thing

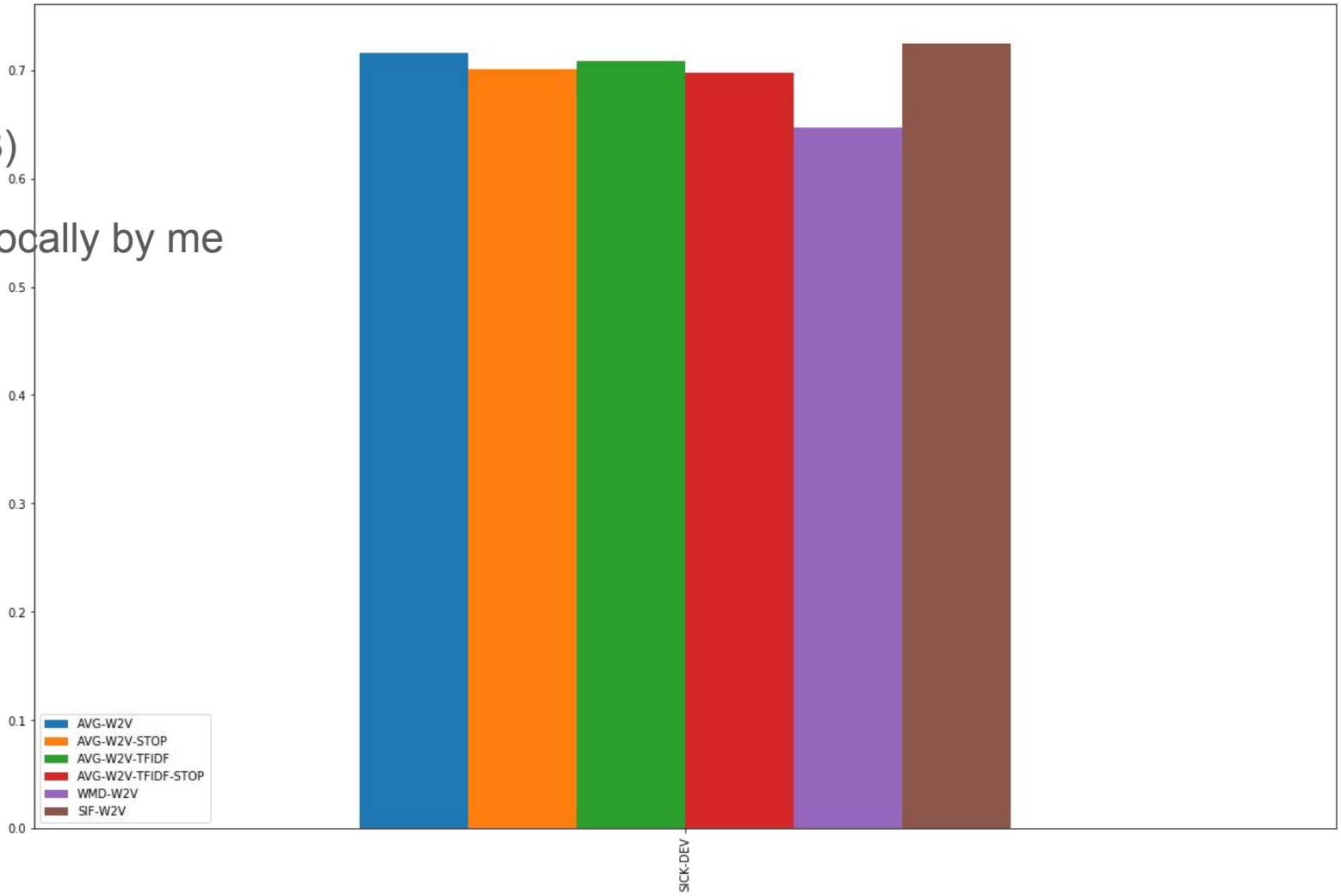
Means: Likely their response is semantically similar to the sentence in the model and the response is correct.

RESULTS

How related are judgements of the machine and human judgements?

Nlptown(2018)

Reproduced locally by me



	score_dist	similarity_sum	D3JOLF	D3JOLR	Accuracy_score	NrRelCorrectO
score_dist	1	0.3506075390...	0.2135199813...	0.2909950575...	0.4433027855...	0.4402623893...
similarity_sum	0.3506075390...		0.4891154336...	0.5308264264...	0.5005812574...	0.4723882212...
D3JOLF	0.2135199813...	0.4891154336...	1	0.71757179114...	0.2944619208...	0.3136710829...
D3JOLR	0.2909950575...	0.5308264264...	0.71757179114...	1	0.3048457105...	0.3471392456...
Accuracy_score	0.4433027855...	0.5005812574...	0.2944619208...	0.3048457105...	1	0.8081139039...
NrRelCorrectO...	0.4402623893...	0.4723882212...	0.3136710829...	0.3471392456...	0.8081139039...	1

Importance of practice testing
Importance of feedback

For students in the room:

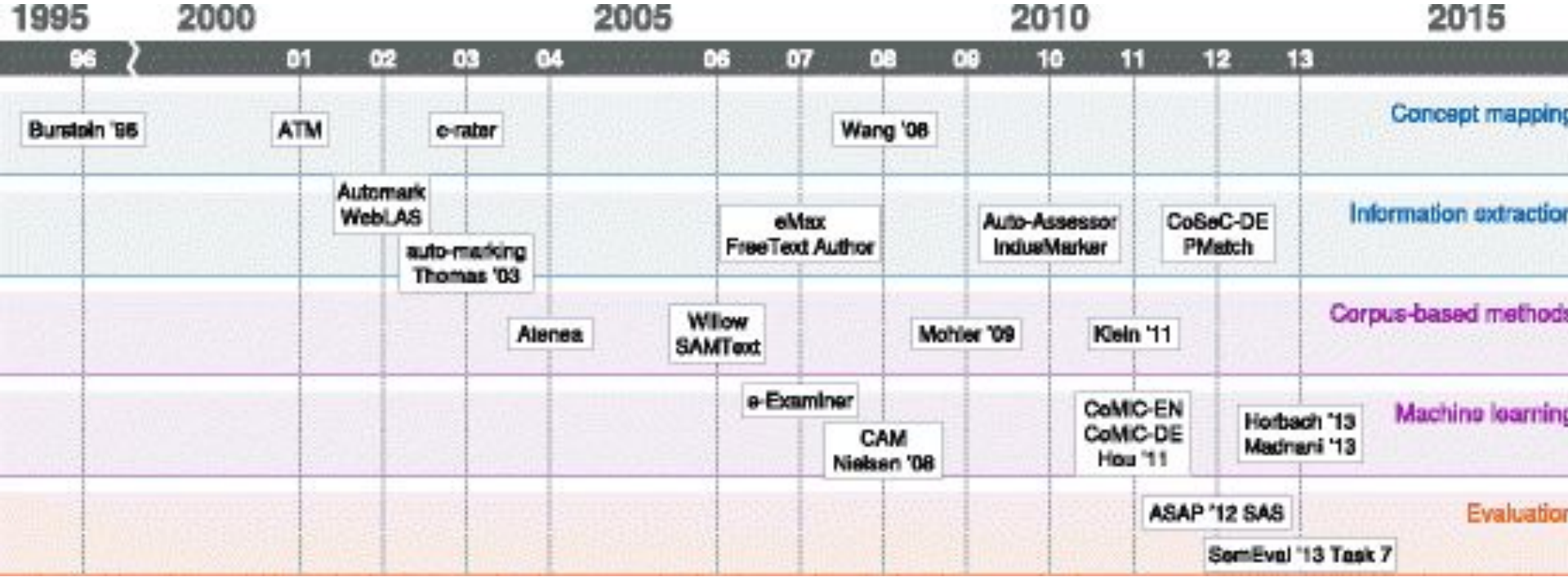
There is plenty of empirical support indicating that as a method of studying practice testing is much more efficient than rereading and summarizing.

Immediate feedback

Alternatives

History of the Field

Automatic short answer grading (ASAG) has so far been a fairly immature and ad-hoc based research field. It uses natural language processing techniques and adjusts them to the specifics of grading short-text responses to question (Burrows et al., 2015)



The most different alternative - UMLS

UMLS - universal medical language system - a system of codified and formalized knowledge for domain of medicine.

It is a language-system approach, where there is a model sentence such as “X causes Y” or “X is part of Y” There are many of the X’s Y’s and a few verbs between them.

Correctness of the answers can be measured easily - the truth value of the sentences is “hard-coded”. It is used in domain specific tasks

Tools I used

pandas

$$y_{it} = \beta' x_{it} + \mu_i + \epsilon_{it}$$

NLPTOWN

NLTK

nlptown@github

https://github.com/karolski/marbel

Repository with:

- Notebooks
- Documentation

Google

glove

All Images Maps Videos Shopping More Settings Tools

About 492.000.000 results (0,60 seconds)

See glove

Sponsored

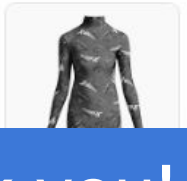
PERFORMANCE HARDLOOPH...
€15,00
Perry Sport

By Periscopix

€69,30
SkiWebShop
21% price drop
By Google

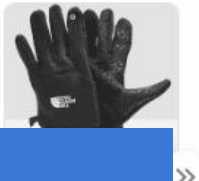
€25,00
Nike Official

By Pricesearcher



€395,00
MatchesFashion

By Pricesearcher



€39,99
Perry Sport

By Periscopix

Thank you!

GloVe: Global Vectors for Word Representation - Stanford NLP

https://nlp.stanford.edu/projects/glove/

GloVe is an unsupervised learning algorithm for obtaining vector representations for words. Training is performed on aggregated global word-word ...

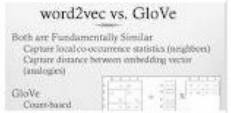
You've visited this page 2 times. Last visit: 1/22/19

People also search for

glove python	glove x_max
glove vs word2vec	how to use glove vectors
glove tutorial	glove python load

GloVe

Machine learning



GloVe, coined from Global Vectors, is a model for distributed word representation. The model is an unsupervised learning algorithm for obtaining vector representations for words. Wikipedia

Feedback

See results about

Glove (Garment)
A glove is a garment covering the whole hand. Gloves usually have separate sheaths or openings ...

Extras:

Google’s sentence encoder

Facebook’s Infersent

Go further:

Googles state-of-art (30.02.2019) free Natural Language processing tool "Bert" in jupyter notebook.link

cross lingual embeddings respository by @nlptown

Kazi, H., Haddawy, P., & Suebnukarn, S. (n.d.). Expanding the Plausible Solution Space for Robustness in an Intelligent Tutoring System. Intelligent Tutoring Systems, 583-592. - article about intelligent tutoring systems - make a feedback hint system helping students learn quickly

Bibiliography on metacognitive monitoring, effects of practice testing on learning and intervention with online tools:

Van Loon, M.H., De Bruin, A.B.H., Van Gog, T., & Van Merriënboer, J.J.G., & Dunlosky, J. (2014). *Can students evaluate their understanding of cause-and-effect relations? The effects of diagram completion on monitoring accuracy.* Acta Psychologica, 151, 143-154.link abstract

Thiede, K. W., Anderson, M. C. M., & Therriault, D. (2003). *Accuracy of metacognitive monitoring affects learning of texts.* Journal of Educational Psychology, 95, 66-73 link

De Bruin, A.B.H., Dunlosky, J., & Cavalcanti. R.B. (2017). *Monitoring and regulation of learning in medical education: The need for predictive cues.* Medical Education, 51, 575-58link

De Bruin, A.B.H, Kok, E.M., Lobbestael, J., & de Grip, A. (2017). *The impact of an online tool for monitoring and regulating learning at university: overconfidence, learning strategy, and personality.* Metacognition and Learning, 12, 21-43.link

De Bruin, A.B.H, & van Merriënboer, J.J. (2017). *Bridging Cognitive Load and Self-Regulated Learning Research: A complementary approach to contemporary issues in educational research.* Learning and Instruction, 51, 1-9.link

Bibliography about methods I used:

- Article about the repository used as a codebase article. It compares different sentence similarity measures:

  Nlptown. (2018, May). Comparing Sentence Similarity Methods. Retrieved from http://nlp.town/blog/sentence-similarity/

- **Word Mover Distance**

  Kusner, M., Sun, Y., Kolkin, N., & Weinberger, K. (2015, June). From word embeddings to document distances. In *International Conference on Machine Learning* (pp. 957-966). paper

- Rong, X. (2014). word2vec parameter learning explained. *arXiv preprint arXiv:1411.2738*. paper
- Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems* (pp. 3111-3119).
- Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. GloVe: Global Vectors for Word Representation. [pdf] [bib] link

- Burrows, S., Gurevych, I. & Stein, B. *The Eras and Trends of Automatic Short Answer Grading,* Int J Artif Intell Educ (2015) 25: 60. https://doi-org.ezproxy.ub.unimaas.nl/10.1007/s40593-014-0026-8

Bibliography about alternative aproaches:

- Alternative language representation: WordNet from Princeton University
- gui application: SEMILAR SEMILAR API comes with various similarity methods based on Wordnet, Latent Semantic Analysis (LSA), Latent Dirichlet Allocation (LDA), BLEU, Meteor, Pointwise Mutual Information (PMI), Dependency based methods, optimized methods based on Quadratic Assignment, etc.
- API serveice cortical.io with similar method

J. Mitchell and M. Lapata, *Composition in Distributional Models of Semantics*, Cognitive Science, vol. 34, no. 8, pp. 1388–1429, Nov. 2010., link

Li, X., & Li, Q. (2015). Calculation of Sentence Semantic Similarity Based on Syntactic Structure. *Mathematical Problems in Engineering*, *2015*, 1-8. doi:10.1155/2015/203475 paper,

Y. Li, D. McLean, Z. A. Bandar, J. D. O'Shea and K. Crockett, *Sentence similarity based on semantic nets and corpus statistics*, in IEEE Transactions on Knowledge and Data Engineering, vol. 18, no. 8, pp. 1138-1150, Aug. 2006.link code

Kazi, Hameedullah & Haddawy, Peter & Suebnukarn, Siriwan. (2012). *Employing UMLS for generating hints in a tutoring system for medical problem-based learning.* Journal of biomedical informatics. 45. 557-65. link *employeeing a predefined ontology systems to produce easily verifiable tasks with a precise hinting system*

Picture links

- https://www.dyclassroom.com/image/topic/python/logo.png
- https://upload.wikimedia.org/wikipedia/commons/thumb/3/38/Jupyter_logo.svg/250px-Jupyter_logo.svg.png
- https://cmg.soton.ac.uk/media/event-images/main-thumb-t-348902-200-dgbemwvkzzlbbhaeukyjituzjzgudxmh.jpeg
- http://nlp.town/assets/img/logo/nlptown_small.png
- https://www.nltk.org/news.html#
-