MT Summer Term 2021 Ex11: Word Embeddings

1. In your own words, why would we like to represent words as vectors (embeddings, sequences of numbers)?

2. In your own words, briefly describe the differences between

- hot-one encodings
- frequency/count-based word embeddings
- prediction based word embeddings

3. In your own words, please describe the difference between "static" and "contextualised" word embeddings.

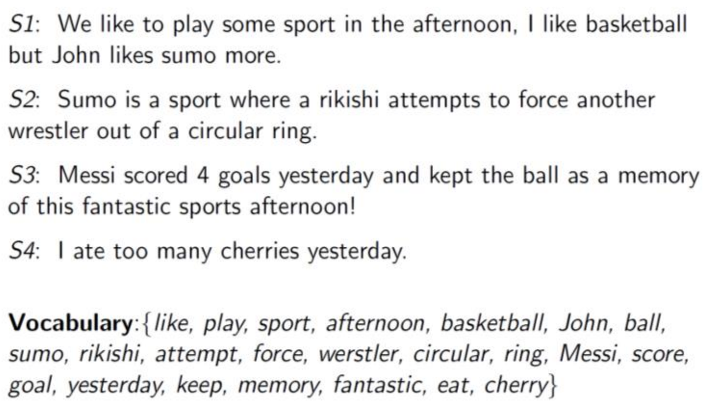
4. Give the following data

I like to play hockey and badminton .

Please construct hot-one encodings for each of the words (including punctuation markers) in the data. Which steps do you have to go through to do this? What's good about the hot one encodings, what's not so good about them?

5. Given two vectors $\mathbf{x}$ and $\mathbf{y}$, please define the Euclidian distance $d(\mathbf{x}, \mathbf{y})$ between $\mathbf{x}$ and $\mathbf{y}$.

6. Given the following data set (from Cristina's slides) and vocabulary (where stop-words have been stripped)



S1: We like to play some sport in the afternoon, I like basketball but John likes sumo more.

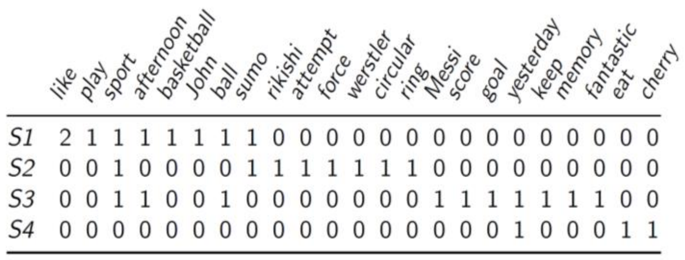
S2: Sumo is a sport where a rikishi attempts to force another wrestler out of a circular ring.

S3: Messi scored 4 goals yesterday and kept the ball as a memory of this fantastic sports afternoon!

S4: I ate too many cherries yesterday.

Vocabulary:{*like, play, sport, afternoon, basketball, John, ball, sumo, rikishi, attempt, force, werstler, circular, ring, Messi, score, goal, yesterday, keep, memory, fantastic, eat, cherry*}

and the following word-document co-occurrence matrix (from Cristina's slides):



	like	play	sport	afternoon	basketball	John	ball	sumo	rikishi	attempt	force	werstler	circular	ring	Messi	score	goal	yesterday	keep	memory	fantastic	eat	cherry
S1	2	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
S2	0	0	1	0	0	0	0	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0
S3	0	0	1	1	0	0	1	0	0	0	0	0	0	0	1	1	1	1	1	1	1	0	0
S4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	1	1

what are the count-based embeddings for *basketball*, *sumo* and *cherry*? What are the Euclidian distances between

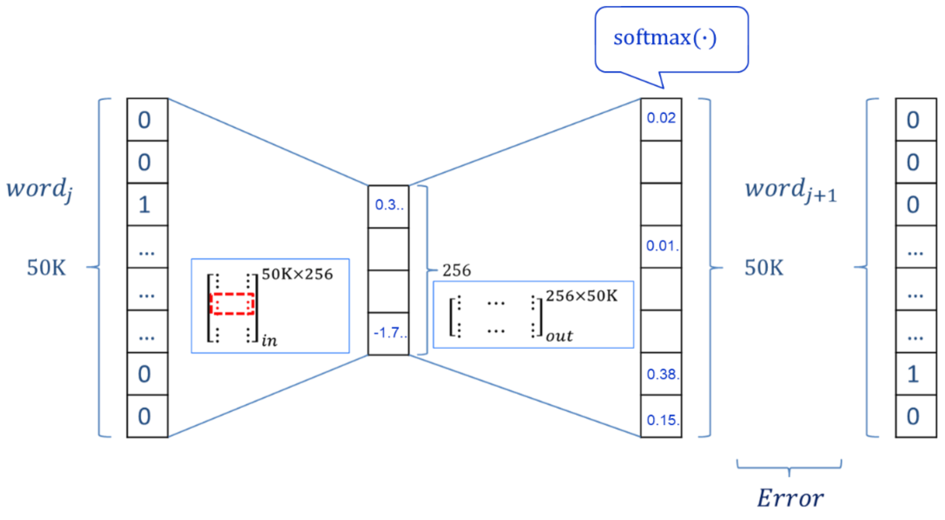
- *basketball*, *sumo*
- *basketball*, *cherry*

- *sumo, cherry*

Is the Euclidian distance between *sumo* and *cherry* the same as that between charry and sumo? Is this generally the case, i.e. is Euclidian distance symmetric?

7. Please define the softmax. Given a vector of real numbers, what does the softmax produce?

8. Please use the following diagram



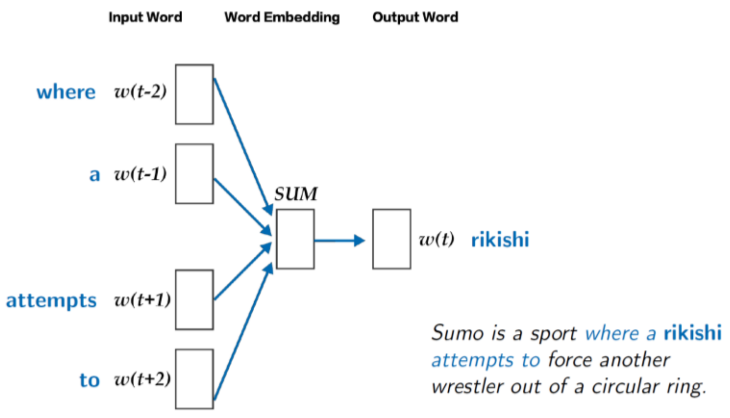
to, in your own words, explain simple prediction-based word embeddings based on a simple bi-gram neural language model. In what sense is the model prediction based. Where are the hot-one vectors? Where are the embedding vectors? Where is the embedding matrix? Where is the projection matrix? In what sense are embedding vectors lower-dimensional dense vectors in continuous vector space? Why do we needs the softmax ? How do we train these embeddings? How do we compute the loss?

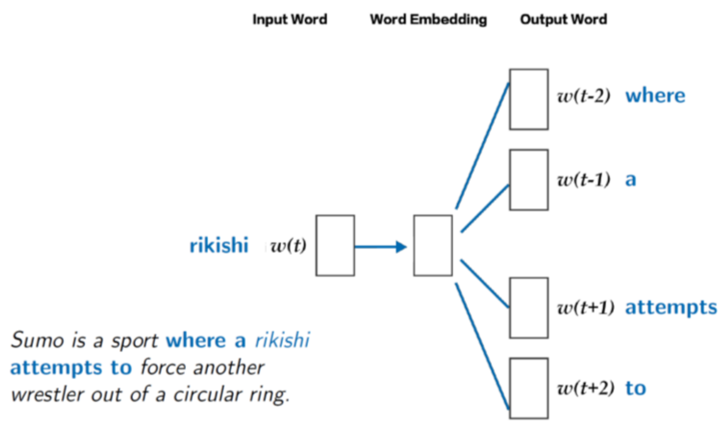
9. In your own words, please explain why the model presented in (8) above is computationally quite expensive. Please, in your own words, explain the following two changes (from Mikolov) to the model that make the model more efficient:

- Replacing the full FFNN by a logistic regression-based classifier and what the classifier classifies;
- Negative Contrastive Estimation

and how the two depend on each other.

10. In your own words, please explain the CBoW and skip-gram models (from Chrsitina's slides) presented by Mikolov (incl. which is which):

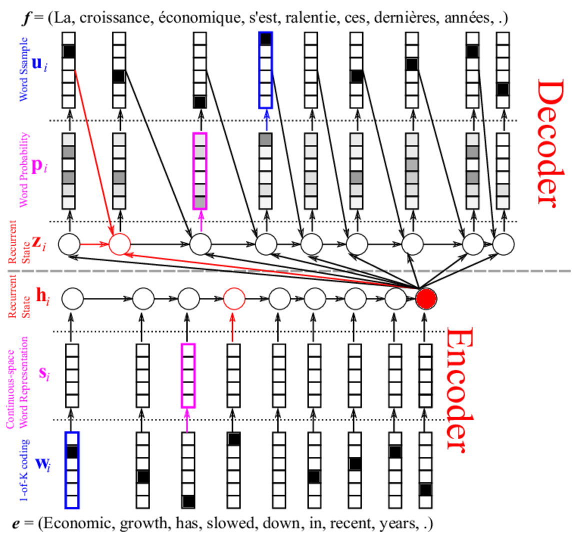




In what sense are they also language model based? In what sense are they static word embeddings (as opposed to contextualized ones)? What are some hyperparameters of these models?

11. What is cool about those prediction and lower-dimensional dense vector space embeddings? Please give some examples (similarity, analogical reasoning) and explain the properties that make such embeddings really nice (part of the "secret sauce" of NN-based approaches to NLP).

11. Please explain where you find hot-one encodings, static word embeddings and contextualized embeddings in the following RNN-based neural machine translation model:


f = (La, croissance, économique, s'est, ralentie, ces, dernières, années, .)
Word Ssample u_i
Word Probability p_i
Recurrent State z_i
Recurrent State h_i
Continuous-space Word Representation s_i
1-of-K coding w_i
Decoder
Encoder
e = (Economic, growth, has, slowed, down, in, recent, years, .)