PROJEKT WYNIKANIE—TEMAT 5

Werydyczność

Zadanie:

DANE: czasownik (cechy + embedding)

WYJŚCIE: sygnatura czasownik (+,-,o / +,-*,o)

Kolumny zbioru danych:

- `GOLD <T, H>` - anotacje par semantycznych
- `GOLD <T1, H>`
- `verb - main semantic class`—główna klasa semantyczna
- `verb` - second semantic class
- `verb` - third semantic class
- **`verb` - czasownik**

Klasyfikator	Zb. treningowy	Zbiór testowy		
	F1	Precision	Recall	F1
`KNeighbors Classifier`	0.720	0.776	0.795	0.782
`RandomForest Classifier`	0.701	0.599	0.697	0.632
`MLPClassifier`	**0.823**	**0.826**	**0.860**	**0.839**
`AdaBoostClassifier`	0.503	0.423	0.648	0.512
`GaussianProcess Classifier`	0.746	0.766	0.795	0.780
`GradientBoosting Classifier`	0.763	0.747	0.795	0.757
`SVC`	0.804	0.811	0.844	0.823

Zbiór danych: zdania podrzędnie złożone (***że***) z anotowanymi cechami czasowników: **2596** zdań, **368** czasowników

Werydyczność: stwierdzenie, czy po użyciu danego czasownika wynika, czy zdanie podrzędne jest prawdziwe (pozytywne i negatywne środowisko)

Opis rozwiązania:

Problem bardzo szerokich danych:

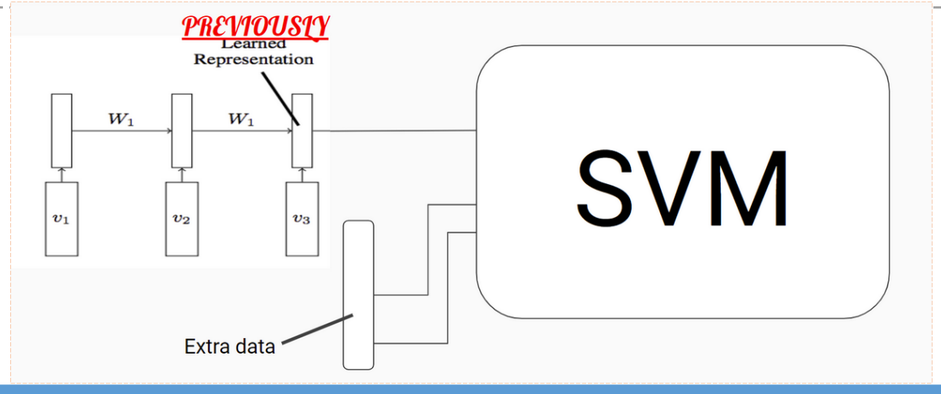
300 wymiarowy embedding $\sim$ 600 unikalnych wierszy po przetworzeniu

Przetestowane podejścia:

- trenowanie modelu na danych bez embeddingów
- trenowanie modelu na danych w postaci samych embeddingów
- trenowanie modelu na danych w embeddingów z dodatkowymi informacjami

Zastosowano Embeddingi typu fastText dla języka polskiego, uśrednianie wyrażeń.

Klasyfikacja wieloetykietowa—zbiór etykiet zbioru stanowią wszystkie kombinacje etykiet obu klas werydyczności.



Autorzy:

Wojciech Bartoszek

Albert Dziugieł

Maciej Gorczyca

Bogdan Jastrzębski