Warsztaty tłumaczenia automatycznego

Sylabus zajęć

Informacje podstawowe

Kierunek studiów Informatyka - przedmioty do wyboru **Specjalność** - **Jednostka organizacyjna** Wydział Matematyki i Informatyki **Poziom studiów** studia drugiego stopnia poinżynierskie **Forma studiów** studia stacjonarne **Profil studiów** profil ogólnoakademicki	**Cykl dydaktyczny** 2023/24 **Kod zajęć** 06INFPWS.4200000S.01049.23 **Języki wykładowe** polski **Obligatoryjność** Fakultatywny **Blok zajęciowy** Przedmioty specjalnościowe

Koordynator zajęć	Artur Nowakowski, Gabriela Pałka
Prowadzący zajęcia	Artur Nowakowski, Gabriela Pałka

Okres	**Forma zajęć / liczba godzin / forma zaliczenia**	**Liczba punktów ECTS**
Semestr letni	• Laboratorium: 30, Zaliczenie z oceną	3

Cele kształcenia dla zajęć

Kod	Cel
C1	Zdobycie wiedzy teoretycznej i praktycznej na temat określania jednostek nazwanych w tekstach pisanych w języku naturalnym.
C2	Nabycie umiejętności pozyskiwania korpusów równoległych.
C3	Nabycie umiejętności tworzenia i praktycznego zastosowania neuronowych systemów tłumaczenia automatycznego.
C4	Zdobycie wiedzy teoretycznej i praktycznej na temat wykorzystania ekstrakcji jednostek nazwanych w tłumaczeniu automatycznym tekstów pisanych w języku naturalnym.
C5	Poznanie metod ewaluacji systemów tłumaczenia automatycznego - poznanie metod ewaluacji rozpoznawania jednostek nazwanych.

Wymagania wstępne

Podstawowa umiejętność programowania w języku Python. Podstawowa umiejętność korzystania z systemu kontroli wersji Git.

Efekty uczenia się dla zajęć

Kod	Efekty uczenia się dla zajęć w zakresie	Efekty uczenia się dla kierunku	Metody weryfikacji osiągnięcia efektów uczenia się dla zajęć
Wiedzy – Student/ka:			
W1	Rozumie pojęcie „jednostka nazwana” i potrafi wskazać jej wystąpienia w tekście.		Projekt, Zadania wykonywane podczas zajęć
Umiejętności – Student/ka:			
U1	Potrafi przygotować korpus danych niezbędny do stworzenia modelu rozpoznawania jednostek nazwanych.		Projekt, Zadania wykonywane podczas zajęć
U2	Potrafi wytrenować model rozpoznawania jednostek nazwanych.		Projekt, Zadania wykonywane podczas zajęć
U3	Umie stworzyć system rozpoznający jednostki nazwane.		Projekt, Zadania wykonywane podczas zajęć
U4	Umie ocenić skuteczność systemu rozpoznawania jednostek nazwanych.		Projekt, Zadania wykonywane podczas zajęć
U5	Rozumie pojęcie korpusu równoległego i potrafi zastosować w praktyce metody jego pozyskania.		Projekt, Zadania wykonywane podczas zajęć
U6	Potrafi wydzielić zbiory danych konieczne do wytrenowania i ewaluacji modelu neuronowego tłumaczenia automatycznego.		Projekt, Zadania wykonywane podczas zajęć
U7	Rozumie pojęcie jednostek podwyrazowych i potrafi zastosować w praktyce algorytm Byte Pair Encoding w celu przetworzenia zbiorów danych.		Projekt, Zadania wykonywane podczas zajęć

Kod	Efekty uczenia się dla zajęć w zakresie	Efekty uczenia się dla kierunku	Metody weryfikacji osiągnięcia efektów uczenia się dla zajęć
U8	Potrafi wytrenować model neuronowego tłumaczenia automatycznego oraz zastosować go do tłumaczenia tekstu.		Projekt, Zadania wykonywane podczas zajęć
U9	Potrafi zastosować rozpoznawanie jednostek nazwanych w celu poprawienia jakości tłumaczenia automatycznego.		Projekt, Zadania wykonywane podczas zajęć
U10	Umie ocenić jakość modelu tłumaczenia automatycznego.		Projekt, Zadania wykonywane podczas zajęć

Treści programowe dla zajęć

Lp.	Treści programowe dla zajęć	Efekty uczenia się dla zajęć	Formy zajęć
1.	Pojęcie i charakterystyka jednostek nazwanych. Różnica pomiędzy rozpoznawaniem jednostek nazwanych a ekstrakcją informacji. Wizualizacja jednostek nazwanych w tekście za pomocą biblioteki spaCy.	W1	Laboratorium
2.	Opracowanie cech dobrego korpusu danych dla zadania rozpoznawania jednostek nazwanych oraz metod jego pozyskiwania. Pozyskanie korpusu danych i przygotowanie go do przetworzenia za pomocą biblioteki flair.	U1	Laboratorium
3.	Implementacja modułu trenującego model rozpoznawania jednostek nazwanych. Wykorzystanie biblioteki flair.	U2	Laboratorium
4.	Implementacja systemu rozpoznającego jednostki nazwane w języku programowania Python 3.	U3	Laboratorium
5.	Charakterystyka metod ewaluacji systemu rozpoznawania jednostek nazwanych. Ewaluacja stworzonego systemu.	U4	Laboratorium
6.	Pojęcie korpusu równoległego. Pozyskanie korpusów równoległych ze zbiorów ogólnodostępnych. Zastosowanie narzędzi ekstrakcji tekstu z dokumentów. Zastosowanie narzędzia hunalign w celu stworzenia korpusu równoległego.	U5	Laboratorium
7.	Wstępne przetwarzanie korpusu. Wydzielenie zbiorów danych uczących i testowych z korpusu przy zachowaniu odpowiednich proporcji.	U6	Laboratorium
8.	Pojęcie jednostek podwyrazowych i algorytm Byte Pair Encoding. Porównanie narzędzi Subword-NMT, fastBPE i SentencePiece. Wytrenowanie modelu SentencePiece i zastosowanie go w celu przetworzenia zbiorów danych.	U7	Laboratorium
9.	Zaznajomienie z narzędziem fairseq. Binaryzacja danych oraz wytrenowanie modelu tłumaczenia automatycznego. Zastosowanie wytrenowanego modelu do tłumaczenia tekstu.	U8	Laboratorium

Lp.	Treści programowe dla zajęć	Efekty uczenia się dla zajęć	Formy zajęć
10.	Pojęcie ograniczeń leksykalnych. Zastosowanie rozpoznawanych jednostek nazwanych jako ograniczeń leksykalnych podczas tłumaczenia automatycznego w celu polepszenia jakości tłumaczenia.	U9	Laboratorium
11.	Ewaluacja ludzka oraz ewaluacja automatyczna z wykorzystaniem metryki BLEU. Ograniczenia metryki BLEU. Wykorzystanie narzędzia GEval do automatycznej ewaluacji stworzonego systemu tłumaczenia.	U10	Laboratorium

Informacje dodatkowe

Forma zajęć	Metody i formy prowadzenia zajęć
Laboratorium	Wykład z prezentacją multimedialną wybranych zagadnień, Metoda laboratoryjna, Metoda warsztatowa, Metoda projektu

Forma zajęć	Warunki zaliczenia zajęć
Laboratorium	Końcowa ocena składa się z następujących elementów: 1. projekt – 20%, 2. zadania wykonywane podczas zajęć – 80%. Skala ocen: 1. bardzo dobry (bdb; 5,0) – od 90% punktów, 2. dobry plus (db plus; 4,5) – od 80% punktów, 3. dobry (db; 4,0) – od 70% punktów, 4. dostateczny plus (dst plus; 3,5) – od 60% punktów, 5. dostateczny (dst; 3,0) – od 50% punktów, 6. niedostateczny (ndst; 2,0) – poniżej 50% punktów.

Literatura

Obowiązkowa

1. Philipp Koehn „Neural Machine Translation”, Cambridge University Press 2020
2. Vaswani et al. „Attention Is All You Need”, NeurIPS 2017
3. Chris Hokamp, Qun Liu „Lexically Constrained Decoding for Sequence Generation Using Grid Beam Search”, ACL 2017
4. Matt Post, David Vilar „Fast Lexically Constrained Decoding with Dynamic Beam Allocation for Neural Machine Translation”, ACL 2018
5. Hu et al. „Improved Lexically Constrained Decoding for Translation and Monolingual Rewriting” , ACL 2019
6. Jing Li, Aixin Sun, Jianglei Han, Chenliang Li „A Survey on Deep Learning for Named Entity Recognition”, IEEE 2020

Nakład pracy studenta i punkty ECTS

Rodzaje zajęć studenta	Średnia liczba godzin* przeznaczonych na zrealizowane rodzaje zajęć
Laboratorium	30
Przygotowanie do zajęć	5

Czytanie wskazanej literatury	5
Przygotowanie projektu	25
Inne	10
Łączny nakład pracy studenta	**Liczba godzin** 75
Liczba punktów ECTS	**ECTS** 3

* godzina (lekcyjna) oznacza 45 minut