# KARTA OPISU PRZEDMIOTU - SYLABUS

Nazwa przedmiotu

Zaawansowane przetwarzanie języka naturalnego

## Przedmiot

Kierunek studiów

Informatyka

Rok/semestr

1/2

Studia w zakresie (specjalność)

Sztuczna inteligencja

Profil studiów

ogólnoakademicki

Poziom studiów

drugiego stopnia

Język oferowanego przedmiotu

polski

Forma studiów

stacjonarne

Wymagalność

obligatoryjny

## Liczba godzin

| Wykład | Laboratoria | Inne (np. online) |
|---|---|---|
| 45 | 15 | |
| Ćwiczenia | Projekty/seminaria | |

## Liczba punktów ECTS

5

## Wykładowcy

Odpowiedzialny za przedmiot/wykładowca:

Mateusz Lango

email: mateusz.lango@cs.put.poznan.pl

tel. 61 665 21 24

Wydział Informatyki i Telekomunikacji

Piotrowo 2, 60-965 Poznań

Odpowiedzialny za przedmiot/wykładowca:

## Wymagania wstępne

Student rozpoczynający ten przedmiot powinien posiadać podstawową wiedzę z rachunku prawdopodobieństwa i statystyki (rozkład normalny, dwumianowy, Bernoulliego, estymacja maksymalnej wiarygodności, estymatory nieobciążone, zgodne, efektywne), a także pogłębioną wiedzę z uczenia maszynowego (klasyfikatory złożone, algorytmy k-NN, Naive Bayes, SVM), a w szczególności uczenia głębokiego (architektury wielowarstwowe, sieci splotowe, wsteczna propagacja błędu). Dodatkowo zakłada się podstawową wiedzę z zakresu przetwarzania tekstu, ekwiwalentną do przedmiotu "Przetwarzanie i wyszukiwanie informacji" lub "Przetwarzanie języka naturalnego" (wyrażenia regularne, stemming, lematyzacja, stopwords, model bag-of-words, miary podobieństwa tekstu).

Student powinien posiadać umiejętność rozwiązywania podstawowych problemów ze statystyki oraz rachunku prawdopodobieństwa, programowania w języku Python wraz z odpowiednią biblioteką do uczenia głębokiego oraz umiejętność pozyskiwania informacji ze wskazanych źródeł.

W zakresie kompetencji społecznych student musi rozumieć, że w informatyce wiedza i umiejętności bardzo szybko stają się przestarzałe, a także prezentować takie postawy jak uczciwość, odpowiedzialność, wytrwałość, ciekawość poznawcza, kreatywność, kultura osobista, szacunek dla innych ludzi.

## Cel przedmiotu

Celem przedmiotu jest zapoznanie studentów z metodologią, zasobami i narzędziami stosowanymi w przetwarzaniu języka naturalnego. Zajęcia skupiają się na omówieniu klasycznych metod statystycznych oraz technik opartych na nowych osiągnięciach głębokiego uczenia maszynowego do problemów takich jak przekład automatyczny, analiza wydźwięku, klasyfikacja tekstów, rozpoznawanie jednostek nazewniczych, analiza składniowa i zależnościowa czy modelowanie tematyczne. Ponadto dodatkowym celem przedmiotu jest kształtowanie umiejętności analizowania modeli statystycznych i uczenia maszynowego pod różnymi względami (złożoność obliczeniowa, rodzaj danych uczących i rozmiar próbki, założenia/ograniczenia modelu, metody wnioskowania) oraz ich wykorzystania do rozwiązywania nietrywialnych problemów dot. zasobów tekstowych.

## Przedmiotowe efekty uczenia się

### Wiedza

Ma zaawansowaną i pogłębioną wiedzę w zakresie konstrukcji systemów informatycznych przetwarzających język naturalny metodami statystycznymi [K2st_W3]

Ma pogłębioną wiedzę o architekturach głębokich sieci neuronowych stosowanych w przetwarzaniu języka naturalnego (w szczególności architektury rekurencyjne i rekursywne) [K2st_W3]

Ma zaawansowaną i pogłębioną wiedzę związaną z wybranymi zagadnieniami, takimi jak: modelowanie języka, analiza składniowa, semantyka dystrybucyjna, wykrywanie jednostek nazewniczych, tłumaczenie maszynowe, systemy konwersacyjne [K2st_W3]

Ma wiedzę o trendach rozwojowych i najistotniejszych nowych osiągnięciach w zakresie przetwarzania języka naturalnego (w tym nowoczesnych architektur głębokiego uczenia maszynowego) [K2st_W4]

Zna zaawansowane metody, techniki i narzędzia stosowane przy budowie translatorów, analizatorów składniowych, systemów dialogowych oraz systemów odpowiadających na pytania [K2st_W6]

Zna zaawansowane metody stosowane przy prowadzeniu prac badawczych w zakresie przetwarzania języka naturalnego [K2st_W6]

### Umiejętności

Potrafi pozyskiwać informacje nt. technik przetwarzania języka naturalnego z literatury oraz innych źródeł (w języku polskim i angielskim), integrować je, dokonywać ich interpretacji i krytycznej oceny, wyciągać wnioski oraz formułować i wyczerpująco uzasadniać opinie [K2st_U1]

Potrafi pozyskiwać odpowiednie zbiory danych do poszczególnych zadań inżynierii lingwistycznej (np. z bazy CLARIN) [K2st_U1]

Potrafi planować i przeprowadzać eksperymenty obliczeniowe na danych tekstowych, interpretować uzyskane wyniki i wyciągać wnioski [K2st_U3]

Potrafi - przy formułowaniu i rozwiązywaniu zadań inżynierskich - integrować wiedzę z różnych obszarów systemów uczących się, inżynierii oprogramowania, przetwarzania języka naturalnego oraz lingwistyki [K2st_U5]

Potrafi ocenić przydatność i możliwość wykorzystania nowych osiągnięć uczenia maszynowego do rozwiązywania problemów dot. przetwarzania tekstu [K2st_U6]

Potrafi określić kierunki dalszego uczenia się i zrealizować proces samokształcenia - w szczególności w zakresie poznawania nowych technik state-of-the-art przetwarzania języka naturalnego [K2st_U16]

Kompetencje społeczne

Rozumie, że w inżynierii lingwistycznej wiedza i umiejętności bardzo szybko stają się przestarzałe [K2st_K1]

Rozumie znaczenie wykorzystywania najnowszej wiedzy z zakresu przetwarzania języka naturalnego w rozwiązywaniu problemów badawczych i praktycznych [K2st_K2]

## Metody weryfikacji efektów uczenia się i kryteria oceny

Efekty uczenia się przedstawione wyżej weryfikowane są w następujący sposób:

a) w zakresie wykładów weryfikowanie założonych efektów kształcenia realizowane jest przez:

- ocenę wiedzy i umiejętności wykazanych na egzaminie pisemnym zawierającym proste zadania problemowe, pytania otwarte oraz pytania w formie testu wielokrotnego wyboru.

- omówienie wyników testu,

b) w zakresie laboratoriów weryfikowanie założonych efektów kształcenia realizowane jest przez:

- ocenianie ciągłe, na każdych zajęciach (odpowiedzi ustne przy tablicy) premiowanie przyrostu umiejętności posługiwania się poznanymi zasadami i metodami oraz narzędziami programistycznymi,

- ocenę i obronę sprawozdań z realizacji zestawów zadań obejmujących zadania obliczeniowe jak i implementacyjne (wymagające wykonania eksperymentów oraz analizy i interpretacji uzyskanych wyników),

- ocenę przygotowanej przez studenta prezentacji omawiającej wybrane zagadnienia z przetwarzania języka naturalnego.

Uzyskiwanie punktów dodatkowych za aktywność podczas zajęć, a szczególnie za:

- omówienia dodatkowych aspektów zagadnienia np. poprzez krótkie prezentacje artykułów naukowych,

- uwagi związane z udoskonaleniem materiałów dydaktycznych,

- wskazywanie trudności percepcyjnych studentów umożliwiające bieżące doskonalenia procesu dydaktycznego.

Zarówno w zakresie wykładów jak i ćwiczeń stosuje się następującą skalę ocen: powyżej 51% punktów - dostateczny, 61% - dostateczny plus, 71% - dobry, 81% - dobry plus, 91% - bardzo dobry.

## Treści programowe

1. Wstęp do przetwarzania języka naturalnego. Język naturalny jako system: próba zdefiniowania języka, podwójna artykulacja systemu językowego, wariatywność języka w ujęciu synchronicznym, relatywizm językowy, teorie uniwersalistyczne. Krótki zarys historyczny badań nad językiem i przetwarzaniem języka oraz ich znaczenie dla rozwoju sztucznej inteligencji. Specyfika i trudności przetwarzania danych tekstowych: pojęcie korpusu, słów spoza słownika, niejednoznaczności. Zastosowania przetwarzania języka naturalnego.

2. Statystyczne modelowanie języka. Modele n-gramowe: estymacja największej wiarygodności, interpolacja liniowa modeli n-gramowych, metoda kubełkowania, metody rozmywania estymat, model back-off Katza oraz ogólny zarys modelu Knesser-Ney'a. Zaawansowane modele języka: model n-gramów klas, grupowanie semantyczne Brown'a, zależności semantyczne w dendogramie grupowania. Ewaluacja modeli języka.

3. Neuronowe modele języka. Modele logarytmiczno-liniowe. Neuronowy autoregresywny model 3-gramowy, model autoregresywny z macierzą zagnieżdżeń. Problem skalowania modeli neuronowych do dużych słowników i rozwiązania: próbkowanie ważone oraz softmax hierarchiczny.

4. Klasyfikacja tekstu. Reprezentacja worka słów (bag-of-words) z reprezentacji wektorowej. Sieci splotowe do klasyfikacji tekstu: warstwa splotu 1D (na znakach i słowach), warstwa pooling-over-time, idea wielu kanałów w kontekście reprezentacji rozproszonej. Studium przypadku: klasyfikacja wydźwięku wypowiedzi: klasyczne podejścia nienadzorowane, model sentymentu Ossgood'a, problem negacji, leksykony sentymentu, przykładowy potok przetwarzania tekstu przy analizie wydźwięku krótkich wypowiedzi użytkowników na przykładzie sieci Twitter.

5. Semantyka. Relacje semantyczne i ich użycie w konstrukcji leksykonów komputerowych: antonimia, homonimia, synonimia, polisemia, homonimia, hiponimia, hiperonimia. Słowosieć. Znaczenie wyrazów a ich własności dystrybucyjne. Macierz słowo-kontekst, pozytywna punktowa informacja wzajemna (PPMI). Reprezentacje rozproszone słów: metody globalne (HAL).

6. Semantyka dystrybucyjna. Reprezentacje rozproszone słów: metody iteracyjne (word2vec), metody dla języków bogatych morfologicznie (FastText). Negatywne próbkowanie. Analogie semantyczne i syntaktyczne wraz z wyjaśnieniem ich powstawania. Problem słów spoza słownika, problem polisemii, problem odwzorowywania uprzedzeń społecznych w reprezentacjach słów.

7. Wstęp do modeli sekwencyjnych. Znakowanie części mowy (PoS) i rozpoznawanie jednostek nazewniczych (NER): definicja problemu, metody ewaluacji i sposoby kodowania. Problem predykcji struktur a problem klasyfikacji.

8. Klasyczne modele do predykcji sekwencji. 3-gramowe ukryte modele Markova (Trigram HMMs), estymacja parametrów modelu, algorytm Viterbiego. Modele Markova o maksymalnej entropii (MEMM). Problem obciążenia etykietą. Inżynieria cech dla problemów NER i PoS.

9. Warunkowe pola losowe do predykcji sekwencji (CRF). Probabilistyczne modele grafowe – skierowane i nieskierowane grafy Markova (powtórzenie): graf rozłożenia na czynniki. Algorytm propagacji przekonania dla łańcuchowych CRF (backward-forward). Trening modeli CRF przy użyciu stochastyczne spadku wzdłuż gradientu.

10. Neuronowa predykcja sekwencji. Rekurencyjne sieci neuronowe (architektura Elmana i Jordana) wykorzystujące reprezentacje rozproszone, przegląd neuronów GRU i LSTM, wykorzystanie warstw CRF, modele dwukierunkowe, algorytm wstecznej propagacji błędu poprzez czas. Sposoby na przyśpieszanie sieci rekurencyjnych.

11. Analiza składniowa. Drzewo wyprowadzania, drzewo zależnościowe, gramatyki bezkontekstowe, problem wieloznaczności, probabilistyczne gramatyki bezkontekstowe (definicja, estymacja, algorytm CKY, forma normalna Chomskiego), wprowadzenie do zleksykalizowanych probabilistycznych gramatyk bezkontekstowych. Zarys rekursywnych sieci neuronowych.

12. Analiza zależnościowa. Metody grafowe i oparte na tranzycjach. Algorytm shift-reduce. Algorytm Chu-Liu-Edmonds'a. Zarys algorytmu Eisnera. Universal Dependencies.

13. Tłumaczenie maszynowe. Źródła trudności związane z automatyzacją przekładu. Piramida Vauquois. Model IBM 1: estymacja parametrów z korpusu zawierającego przypisania wyrażeń do ich tłumaczenia, estymacja parametrów z korpusu równoległego.

14. Problem dopasowania słów. Algorytm maksymalizacji oczekiwań dla modelu IBM 1. Wstęp do tłumaczenia frazowego. Ewaluacja systemów tłumaczenia maszynowego (ocena ekspercka i automatyczna - BLEU).

15. Neuronowe metody tłumaczenia maszynowego. Podejścia typu enkoder/decoder, podejścia z uwagą. Reprezentacje rozproszone niezależne od języka. Współdzielenie enkodera, technika backtranslation. Bezpośrednia optymalizacja BLEU: uczenie minimalnego ryzyka.

16. Transfer wiedzy w inżynierii lingwistycznej. Wykorzystanie wiedzy z modeli języka do klasyfikacji tekstu i innych zadań: metoda ULMFit. Kontekstowe zagnieżdżenia słów na przykładzie ELMO.

17. Transfer lingwistyczny: metody mapowania zagnieżdżeń słów w sposób nadzorowany i nienadzorowany. Ortogonalny problem Procrustesa i jego rozwiązanie poprzez dekompozycję SVD macierzy. Problem inicjalizacji macierzy tłumaczeń.

18. Architektura transformer. Problemy głębokich sieci rekurencyjnych z neuronami GRU/LSTM. Quasi-rekurencyjne sieci neuronowe. Idea wielogłowicowej uwagi. Zagnieżdżenia pozycyjne.

19. Wielozadaniowe modele Inżynierii Lingwistycznej. Umówienie publicznie dostępnych systemów opartych na architekturze transformer: BERT, Universal Sentence Encoder, GPT-3 i podobne. Zarys architektury Reformer. Wielojęzykowe modele BERT oraz model dla języka polskiego HerBERT. Możliwości i ograniczenia nowoczesnych modeli języka.

20. Przegląd wybranych zagadnień dot. przetwarzania języka naturalnego (wybór wg zainteresowań studentów): metody text-to-speech, techniki rozpoznawania mowy (ASR), budowanie grafów wiedzy z tekstów, odpowiadanie na pytania, wyszukiwanie informacji (modele DSSM), systemy dialogowe, modelowanie tematyczne.

W ramach laboratoriów studenci wykonują eksperymenty obliczeniowe oraz implementują modele omówione na wykładzie. Ponadto studenci prezentują wybrane artykuły naukowe na zajęciach w formie krótkich prezentacji.

## Metody dydaktyczne

1. Wykład: prezentacja multimedialna, ilustrowana przykładami podawanymi na tablicy, ćwiczenia praktyczne (m.in. obliczeniowe na tablicy).

2. Laboratoria: zadania obliczeniowe i implementacyjne (np. w formie Jupyter Notebook), dyskusja zagadnień i rozwiązań.

## Literatura

### Podstawowa

1. Jurafsky D., Martin J.H.: Speech and Language Processing, III edycja, Pearson/Prentice Hall, 2018 (dostęp online: https://web.stanford.edu/~jurafsky/slp3/)

2. Li Deng, Yang Liu: Deep Learning in Natural Language Processing. Springer, 2018 (dostęp poprzez eZasoby biblioteki PP)

### Uzupełniająca

1. Mykowiecka, A: Inżynieria lingwistyczna: komputerowe przetwarzanie tekstów w języku naturalnym, Wydawnictwo PJWSTK, 2007.

2. Yoav Goldberg, Neural Network Methods in Natural Language Processing (Synthesis Lectures on Human Language Technologies), Morgan & Claypool Publishers, 2017.

3. Goodfellow I., Yoshua B., Courville A.: Deep Learning. Systemy uczące się., PWN, 2018

4. Lango M., Brzeziński D., Stefanowski J.: PUT at SemEval-2016 Task 4: The ABC of Twitter Sentiment Analysis, Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval 2016), 2016

## Bilans nakładu pracy przeciętnego studenta

| | Godzin | ECTS |
|---|---|---|
| Łączny nakład pracy | 125 | 5,0 |
| Zajęcia wymagające bezpośredniego kontaktu z nauczycielem | 60 | 3,0 |
| Praca własna studenta (studia literaturowe, przygotowanie do egzaminu, wykonanie zadań domowych)[1] | 65 | 2,0 |

[1] niepotrzebne skreślić lub dopisać inne czynności