Zasady zaliczenia przedmiotu *Inżynieria języka naturalnego*

Maciej Piasecki

4 października 2016

1 Zasady zaliczenia wykładu

1. Podstawą zaliczenia będą wyniki sumaryczne uzyskane z dwóch kolokwiów.
2. Dokładne terminy kolokwiów zostaną ustalone ze studentami na początku semestru, ale pierwsze będzie nie wcześniej niż na 7 wykładzie, a drugie nie później niż na ostatnim wykładzie.
3. Z każdego kolokwium wynik punktowy będzie przeliczany na wynik procentowy.
4. Ocena końcowa zależy wyłącznie od sumy wyników procentowych z obu kolokwiów liczby punktów uzyskanej na kolokwium. Ocena 3.0 jest od sumy równej 60, każda kolejna jest od sumy o 20 większej.
5. Ocena celująca może być przyznana przy bardzo wysokiej sumie i twórczym podejściu do odpowiedzi na pytania problemowe.

2 Zasady zaliczenia projektu

1. Realizowane tematy dotyczą przetwarzania tekstów lub zapisów mowy w języku polskim.
2. Zajęcia będą realizowane przez grupy 2-3 osobowe.
3. W niektórych przypadkach prowadzący może zgodzić się na indywidualną realizację projektu.
4. Projekt ma być wykonywany systematycznie przez cały semestr zgodnie z harmonogram uzgodnionym z prowadzącym i dostarczonym na piśmie najpóźniej na 3 zajęciach.
5. Na każdych zajęciach prowadzący odnotowuje postęp prac. Uchybienia w systematycznej realizacji mogą skutkować obniżeniem oceny końcowej.
6. Zgodnie z etapami w harmonogramie studenci oddają krótkie notki z opisem postępu prac.
7. Rozliczenie projektu musi nastąpić nie później niż na ostatnich zajęciach.

8. Ocena końcowa zostanie ustalona z uwzględnieniem wyników dostarczonych do regulaminowego końca ostatnich zajęć z uwzględnieniem odnotowanego obrazu systematyczności pracy.

Proponowany ramowy harmonogram realizacji projektu jest następujący (w nawiasach podano przewidywany czas trwania etapu w tygodniach:

- wybór tematu oraz jego szczegółowe doprecyzowanie (1)
- przegląd literatury oraz zapoznanie się z niezbędnymi podstawami teoretycznymi (2)
- wybór metod do implementacji (1)
- wybór niezbędnych narzędzi i komponentów programistycznych (1)
- instalacja wybranych komponentów programistycznych, zapoznanie się z ich wykorzystaniem, zintegrowanie ze stosowanym środowiskiem developerskim (1)
- implementacja (3)
- zgromadzenie niezbędnych danych do badań i eksperymentów (2)
- przeprowadzenie eksperymentów, optymalizacja parametrów metod (2)
- przygotowanie raportu końcowego (1)

W przypadku konkretnych projektów czas trwania faz może być nieco inny.

3 Tematy do wyboru

1. **Porównanie cech bi-gramowych modeli językowych tworzonych z wykorzystaniem różnych metod wygładzania**

   Budowa modelu statystycznego języka polega na wyznaczeniu estymowanego prawdopodobieństwa wystąpienie słowa w określonym kontekście, zwykle jednostronnym. Prawdopodobieństwa te wyznacza się na podstawie częstości występowania n-elementowych sekwencji słów wystarczająco obszernej próbce języka zwanej korpusem. Jednak przy wielkiej liczbie wyrazów i ich form fleksyjnych występujących w języku nie sposób zgromadzić korpusu, w którym wystąpią wszystkie możliwe sekwencje słów, stąd stosując zwykły estymator częstościowy dla większości n-gramów estymowane prawdopodobieństwo byłoby zerowe. Dlatego stosuje się różne metody pozwalające na przypisanie niezerowych prawdopodobieństw również tym n-gramom, które nie wystąpiły w korpusie. Zabieg ten zwany jest wygładzaniem. W dostępnej literaturze opisano wiele metod wygładzania. Otrzymany model można ocenić za pomocą parametru zwanego perpleksją skrośną.

   Celem projektu jest oprogramowanie wybranych metod tworzenie statystycznych modeli języka z użyciem wygładzania oraz ocena uzyskanych modeli. Realizacja zadania obejmuje:

- zapoznanie się z typowymi metodami wygładzania opisanymi w literaturze,
- pozyskanie korpusów tekstów w języku polskim do wykonania eksperymentów, najlepiej sprofilowanych dziedzinowo,
- zapoznanie się z typowym formatem zapisu modeli językowych (np. format DARPA stosowany w HTK i innych popularnych narzędziach ASR),
- implementacja metod tworzenia modelu, jego zapisu w standardowym formacie oraz procedury oceny uzyskanych modeli,
- wykonanie badań dla zgromadzonych korpusów.

2. **Korekta błędów w tekście polegających na mylnym złączaniu/rozdzielaniu wyrazów**

   W procesie automatycznego rozpoznawania tekstu drukowanego bardzo często występują błędy polegające na mylnym rozdzielaniu lub złączaniu wyrazów, w szczególności jeśli system OCR nie jest wspomagany słownikiem (np. program ABBY's PDF Transformer wykorzystywany do tekstów w języku polskim). Celem projektu jest opracowanie metody dokonującej korekcji tego rodzaju błędów. Zakładamy, że dostępny jest bi-gramowy model języka oraz słownik który jest częścią tego modelu (sekcja unigramów jest w istocie takim słownikiem). Jedna z możliwych metod może być następująca. Dla słów które nie występują w słowniku badamy możliwości ich rozdzielenia lub połączenia z sąsiednimi słowami. Każdy wariant oceniamy za pomocą perpleksji skrośnej wyznaczonej na podstawie modelu językowego. Ostatecznie wybieramy ten wariant, który daje najlepsza ocenę. Oczywiście mile widziane są inne własne pomysły.

3. **Automatyczne wstawianie znaków interpunkcyjnych do tekstu.**

   Sformułowanie nie budzi chyba wątpliwości. Na wejściu mamy tekst składający się ze słów ale pozbawiony znaków interpunkcyjnych z wyjątkiem kropek kończących zdania. Celem jest wstawienie znaków interpunkcyjnych (kropki, średniki myślniki) stosownie do zasad interpunkcji w języku polskim. W ramach projektu należy opracować zbiór zasad wstawiania znaków interpunkcyjnych, zaimplementować je w postaci programu oraz zbadać skuteczność opracowanej metody eksperymentalnie.

4. **Łączenie złamanych słów i rozdzielonych zdań.**

   Podczas skanowania tekstu i rozpoznawania za pomocą OCR, a nawet podczas generowania PDF-ów dochodzi często do dzielenia słów na dwie części w miejscu przenoszenia do nowej linii, np. „dokonuje się za-

   zwyczaj za sprawą pokolenia". Często również na przejściach pomiędzy stronami zdanie zostaje przecięte przez numera strony, stopkę lub przypisy, np.
   „Nie ma nawet takiego pojęcia zbior-

   czego jak Młoda Polska, którym nakryto w końcu urozmaicone

literatury przełomu wieków." (Terasa Walas, Współczesna literatura polska – między empirią a konceptualizacją. *Teksty Drugie*, Vol. 1, Num. 6, 1990)
Celem zadania jest opracowanie programu, który umożliwi automatyczną korektę obu problemów w tekstach tego typu.

5. **Program do wydobywania kolokacji form podstawowych (lematów) z wielkich korpusów polskich tekstów.**
Kolokacją nazywamy silne związki frazeologiczne n słów, które są używane znacząco często razem jako jedna jednostka wypowiedzi językowe i których znaczenie często nie jest prostą pochodną znaczeń wyrazów składowych. Przykładem kolokacji są takie związki wyrazowe jak: *czerwona kartka* czy *zgniły kompromis*. Kolokacje można wykrywać metodami analizy statystycznej. Znacznie lepiej dopracowane jest wydobywanie kolokacji dwuelementowych w oparciu o miary asocjacyjne, ale również istnieją metody dla kolokacji n-elementowych. Lemat (podstawowa forma morfologiczna) to forma hasłowa wyrazu, czyli taka jak używana w słowniku. Tekst można sprowadzić do lematów, np. dla języka polskiego za pomocą tagera WCRFT (`http://ws.clarin-pl.eu`).

   Wiele metod wydobywania kolokacji zostało zaimplementowanych w ramach systemu *MeWeX* (CLARIN-PL). W tym również miary złożone oparte na połączeniu wielu miar asocjacyjnych za pomocą sieci neuronowej czy też kombinacji liniowej dostrajanej metodami ewolucyjnymi. W ramach zadani można również podjąć próbę rozszerzenia MeWeXa o nowe miary lub o nowe sposoby łączenia znanych miar prostych. Jako zbiór wzorcowy można wykorzystać 55 tys. kolokacji sprawdzonych ręcznie i opisanych w Słowosieci.

6. **Program do wydobywania kolokacji opisanych strukturalnie z wielkich korpusów tekstu w oparciu o wydobyte wcześniej kolokacje form podstawowych (realizacja tylko w powiązaniu z tematem powyżej).**

   Kolokacje rozpoznane na poziomie lematów mogą zawierać wiele nadmiarowych, niepoprawnych kolokacji, ponieważ nie uwzględniają składniowych powiązań słów. Aby odfiltrować tylko poprawne składniowo kolokacje można zastosować filtry na poziomie struktury składniowo semantycznej wyrażeń kandydujących, a nawet semantycznej. Na potrzeby opisu wielowyrazowych haseł (dokładnie jednostek leksykalnych) w Słowosieci zostało opracowanych szereg wzorców fraz wyrażonych w języku WCCL. Zadaniem byłoby zastosowanie tych wzorców do filtrowania kolokacji wykrywanych na poziomie ciągów lematów (np. otrzymywanych z *MeWeX*), przebadanie ich skuteczności na podstawie danych wzorcowych ze Słowosieci, a następie ich ewentualnym rozszerzeniu.

Do filtrowania semantycznego można użyć modeli podobieństwa semantycznego słów generowanych z dużych zbiorów tekstu za pomocą pakietu *Word2Vec*. Gotowe modele można otrzymać z CLARIN-PL.

7. **Wydobywanie słów kluczowych z niewielkich dokumentów i fragmentów dokumentów.**

   W eksploracji dużych zbiorów dokumentów tekstowych bardzo pomocnym jest dostęp do słów kluczowych opisujących danych dokument. Bardzo często jednak słowa kluczowe nie są dostępne lub nie odpowiadają w pełni treści dokumentu. W literaturze zaproponowano szereg algorytmów pozwalających na wygenerowanie słów kluczowych na podstawie płytkiej, statystycznej analizy treści dokumentów.

   W ramach projektu należy przeprowadzić eksperymenty z kilkoma algorytmami wydobywania słów kluczowych. Słowa kluczowe powinny semantycznie reprezentować dany dokument lub jego fragment. Zakładamy tu wykorzystanie wszelkiej dostępnej wiedzy lingwistycznej o tekście dokumentu i jego strukturze.

   Do budowy programu będzie można wykorzystać zbiór opisany ręcznie przygotowany przez Grupę Naukową G4.19 lub dostępne zbiory prac naukowych lub abstraktów naukowych z przypisanymi słowami kluczowymi.

8. **Wykrywanie nagłówków w dokumentach tekstowych.**

   Rozważamy tu różnego typy dokumenty, albo czysto tekstowe, zapisane w HTML-u, wydobyte z PDF-ów itd. Ważne jest, że pomimo pewnych informacji strukturalnych jak podział na wiersze, puste linie, znaczniki itd. nie mamy bezpośredniej informacji co jest nagłówkiem akapitu, tytułem, śródtytułem itd., a co jest opisywaną treścią.

   Technika dowolna, choć na pewno narzędzia językowe się przydadzą.

9. **Program do rozpoznawania opisów morfologicznych słów oparty na maszynowym uczeniu się.**

   Celem jest zbudowanie programu do rozpoznawania opisu morfologicznego, np. według modelu Narodowego Korpusu Języka Polskiego, dla słów spoza słownika zawartego w analizatorze morfologicznym *Morfeusz*.

   Dane uczące można wydobyć ze słowników Morfeusza lub wygenerować za pomocą Morfeusza i dużego zbioru tekstów. Istnieje wiele różnych podejść do predykcji morfologicznej: od konstrukcji automatów do różnych metod maszynowego uczenia się.

   Można też zbudować program uwzględniający kontekst wystąpienia słowa, np. por. prace na temat konstrukcji tagera *WCRFT*.

10. **Analiza zastosowania różnych klasyfikatorów do problemu nadzorowanego i nienadzorowanego WSD dla języka polskiego.**
    WSD = *Word Sense Disambiguation* – problem wyboru właściwego znaczenia słowa wieloznacznego w kontekście jego użycia. Problem ujednoznaczniania znaczeń polega na wskazaniu, które znaczenie danego słowa zostało aktywowane w danym kontekście. Dla przykładu w zdaniu *Zamek to budowla obronna (. . .)* zostało aktywowane znaczenie zamku dotyczące

budowli a nie suwaka. Problem ten można sprowadzić do problemu wieloklasowej klasyfikacji. Z punktu widzenia problemu klasyfikacji bardzo istotnym zagadnieniem jest selekcja cech. Celem jest przebadanie wpływu różnych klasyfikatorów i różnego doboru cech na uzyskiwane wyniki. Dane uczące są dostępne z CLARIN-PL. Do wydobycia cech można wykorzystać system *Fextor* lub uproszczony *Fector2* ze systemu do stylometrii o nazwie *WebSty*.

11. **Zastosowania metod nadzorowanej klasyfikacji do rozpoznawania autorstwa lub stylu tekstu**

    W oparciu o opis dokumentów tekstowych za pomocą różnych cech można zbudować klasyfikator, którego celem będzie ustalenie czy dwa dokumenty zostały napisane przez tego samego autora lub w tym samym stylu literackim. Zadanie może być realizowane jako rozszerzenie otwartego systemu *WebSty* z CLARIN-PL.

12. **Klasyfikacja zapytań dla systemu odpowiedzi na pytania dla języka polskiego.**
    W systemie odpowiedzi na pytania wyrażone w języku naturalnym bardzo ważne jest rozpoznanie typu pytania i wydobycie z pytania kluczowych terminów oraz typu oczekiwanej odpowiedzi. Celem jest zbudowanie klasyfikatora do rozpoznania typu pytania i wyznaczenia tych jego elementów, które reprezentują najlepiej klasę semantyczną przedmiotu pytania.

    Zbiory treningowo-testowe, np. korpus *CzyWiesz*, można uzyskać z CLARIN-PL.

13. **Analiza zastosowania różnych klasyfikatorów jako podstawy działania chunkera dla języka polskiego, w tym CRF.**
    Chunker to analizator składniowy produkujący uproszczoną reprezentację składniową zdania w postaci podziału zdania na główne bloki. Celem jest przetestowanie wpływu różnych klasyfikatorów na wyniki chunkera. Podstawą będzie duży zbiór zdań ręcznie opisanych w ramach *Korpusu Politechniki Wrocławskiej*. Można wykorzystać i rozszerzyć istniejący kod chunkera o nazwie *Iobber* zbudowany w G4.19.

14. **Program uczący się rozpoznawania relacji semantycznych pomiędzy bytami nazwanymi na podstawie anotowanego korpusu.**

    Automatyczne wykrywanie w tekście relacji typu: *zlokalizowany w* lub *pracuje w* jest bardzo ważnym elementem wydobywania informacji. *Byt nazwany* to wyrażenie językowe odnoszące się do obiektu z jednego z predefiniowanych typów. Celem jest w oparciu o metody maszynowego uczenia się i duży zbiór instancji relacji oznaczonych ręcznie w tekście zbudowanie programu do wykrywania i klasyfikowania relacji. Zostanie wykorzystany system *Inforex* (skonstruowany w ramach G4.19) implementujący różne techniki wydobywania informacji z polskich tekstów. Dodatkowe informacje:

    - dane testowe dostępne z korpusu InfiKorp w formacie *ccl*,
    - istnieje możliwość opracowania własnego korpusu testowego przy pomocy systemu *Inforex*,

- można oprzeć rozwiązanie na reimplementacji metody Rapier, (Mary Elaine Cali?. Relational learning techniques for natural language information extraction. Doctor of philosophy, The University of Texas at Austin),
- lub na reimplementacji metod oparty na funkcjach jądrowych (kernels), (Razvan Constantin Bunescu. Learning for information extraction: from named entity recognition and disambiguation to relation extraction. Ph.d., The University of Texas at Austin, 2007.).

15. **Rozpoznawanie relacji semantycznych opisywanych w Słowosieci według metody Girju i inni (2006).**

    Należy dokonać re-implementacji jednego ze znanych algorytmów do wydobywania z korpusu tekstów par słów powiązanych określoną relacją, np. hiperonimii (taksonomiczną), meronimii (część-całość) raz z różnymi podtypami, relacji pomiędzy rzeczownikami i przymiotnikami itd.

    Przykłady podejść:

    - hiperonimia dystrybucyjne podejście: https://www.aclweb.org/anthology/P/P14/P14-1113.xhtml
    - hiperonimia wzorce: http://www.aclweb.org/anthology/Y10-1038
    - dla meronimii: Girju, R.; Badulescu, A. & Moldovan, D. Automatic Discovery of Part-Whole Relations. Computational Linguistics, 2006, 32, 83-135

    Celem jest wydobycie par słów na potrzeby na potrzeby półautomatycznego rozszerzania Słowosieci.

16. **System dialogowy do systemu odpowiedzi na pytania.** Celem jest zaprojektowanie i konstrukcja prostego systemu dialogowego jako interfejsu użytkownika do systemu odpowiadającego na pytania w języku naturalnym. Można użyć dowolnego gotowego szkieletowego systemu do konstrukcji *chatterboota*. Istnieje możliwość wykorzystania regułowego modułu do analizy pytań bazującego na regułach w języku *WCCL* skonstruowanego w ramach G4.19. Moduł umożliwia częściowe dopasowanie do zdefiniowanych reguł i transformację praktycznie do dowolnej reprezentacji (wg. określonego szablonu).

17. **Wydobywanie słowników dwujęzycznych z korpusów dwujęzycznych.**
    Podstawą będą dostępne dwujęzyczne i wielojęzyczne korpusy tekstu, które zawierają wierne tłumaczenia dokumentów tekstowych, np. unijnych. Celem będzie zastosowanie dowolnego narzędzia do odniesienia tekstów na poziomie zdań i słów, a następnie wydobycie par słów, które są potencjalnie swoimi tłumaczeniami. Można wykorzystać istniejące dostępne narzędzia.

18. **Program do automatycznego rzutowania Słowosieci na Princeton WordNet.**

Celem jest zaimplementowanie wybranej metody rzutowania jednego wordnetu (elektronicznego tezaurusa) na drugi. Zostanie zastosowany słownik pol.-ang. zebrany ze źródeł w ramach prac G4.19.

19. **Zestaw narzędzi do semantycznej diagnostyki struktury Słowosieci.**

    Słowosieć, podobnie jak inne wordnety, może zawierać szereg błędów w swojej strukturze. Celem jest zebranie z literatury opisów procedur diagnostycznych opartych o dane pozyskane z różnych źródeł: słowników, ontologii i encyklopedii, jak również o funkcję podobieństwa wydobytą z korpusu tekstów. Następnie zastosowanie wybranej metody do oceny Słowosieci 3.0.

20. **Klasyfikacja tematyczna tekstu polskiego na podstawie Słowosieci.**

    Analizując położenia poszczególnych słów tekstu w strukturze Słowosieci można spróbować określić tematykę tekstu, np. względem zbioru wzorcowego. Można rozważyć zastosowanie systemu WoSeDon do ujednoznaczniania znaczeń słów (dostępny również jako web service: `http://ws.clarin-pl.eu`), aby określić znaczenia w jakich zostały użyte niektóre słowa w tekście.

21. **Metody grupowania tekstu polskiego w oparciu o metody inżynierii języka naturalnego.**

    Celem jest zastosowanie znanych algorytmów grupowania i dostępnych narzędzi do analizy słów i struktury polskiego tekstu do podziału kolekcji dokumentów na podzbiory spójne tematycznie.

22. **Metody klasyfikacji polskich tekstów w oparciu o metody inżynierii języka naturalnego.**

    Celem jest zastosowanie dostępnych narzędzi do analizy języka polskiego do poprawy wyników automatycznej klasyfikacji tematycznej polskich tekstów.

23. **Tworzenie mapy dokumentów w języku polskim w oparciu o metody inżynierii języka naturalnego.**

    Celem jest zastosowanie dostępnych narzędzi do analizy języka polskiego w ramach konstrukcji mapy semantycznej (np. w oparciu o algorytm SOM) kolekcji polskich dokumentów.

24. **Automatyczne rzutowania Słowosieci na Wikipedię i *vice versa***

    Wiele słów występujących w Słowosieci jest również opisanych w Wikipedii. Często opisane są te same znaczenia, często jednak podział na znaczenia w obu zasobach jest różny. Istnieje wiele metod łączenia wordnetów z encyklopediami typu Wikipedia. Celem jest wybranie jednej z metod i jej zastosowanie do automatycznej konstrukcji powiązania pomiędzy Słowosiecią a Wikipedią.

25. **Analiza polaryzacji nastawienia emocjonalnego tekstów** Celem jest określenie polaryzacji nastawienia emocjonalnego tekstów (tzw. sentymentu) w oparciu o opis nastawienia znaczeń leksykalnych w Słowosieci i/lub model nauczony na podstawie komentarzy użytkowników. Korpusy komentarzy dostępne są z GRupy Naukowej G4.19.

26. **Automatyczne rozszerzenie opisu polaryzacji nastawienia emocjonalnego w Słowosieci**

    31 tys. znaczeń leksykalnych w Słowosieci zostało opisanych ręcznie pod względem polaryzacji nastawienia emocjonalnego oraz podstawowych emocji. Celem zadania jest zastosowanie jednego ze znanych algorytmów, który rozszerzy ten opis automatycznie na możliwe dużą część Słowosieci. Algorytmy tego typu wykorzystują graf relacji Słowosieci oraz informację o typach relacji do inteligentnego rozszerzania opisu.