Akademia Górniczo - Hutnicza
im. Stanisława Staszica
w Krakowie

Wydział Elektrotechniki, Automatyki i Informatyki Biomedycznej

Autoreferat rozprawy doktorskiej

Modelowanie języka naturalnego (polskiego) dla potrzeb budowy systemu rozpoznawania mowy klasy LVCSR

Natural language modeling of Polish language for purposes of construction Large Vocabulary Continuous Speech Recognition system

---

mgr inż. Leszek Gajecki

Promotor:
prof. dr hab inż. Ryszard Tadeusiewicz

Kraków 2013

Streszczenie

Przedmiotem tej rozprawy jest modelowanie języka naturalnego (polskiego) dla potrzeb budowy systemu rozpoznawania mowy ciągłej z zastosowaniem dużego słownika LVCSR.

Aby osiągnąć ten cel wykazano tezy:

1. Jest możliwe wykonanie modelu językowego dla języka polskiego z użyciem sieci neuronowych - sieci samoorganizującej się mapy Kohonena.
2. Model języka naturalnego oparty na automatycznej ekstrakcji reguł pozwala uzyskać poprawę jakości rozpoznawania mowy polskiej w systemie klasy LVCSR w stosunku do szeroko stosowanego modelu bigramowego.

Model n-gramowy typowo stosowany w badaniach w niektórych językach (np. angielskim) jest rozwiązaniem wygodnym i wydajnym, jednakże sztywno uwzględnia kolejność wyrazów. Język polski potrzebuje takiej klasy modeli językowych, które uwzględniają swobodny (w pewnym zakresie) szyk zdania.

Tezy wykazano poprzez analizę uwarunkowań pracy modeli językowych, analizę znanych modeli statystycznych, gramatyk formalnych, wskazanie najważniejszych elementów gramatyk języka polskiego. Następnie zaproponowano model gramatyki powierzchniowej oparty o proste reguły językowe. Obszerną część pracy poświęcono autorskim modelom opartym o sieci neuronowe, które spełniają uwarunkowania szyku swobodnego.

Dodatkowo reguły językowe w modelach bazujących na sieciach neuronowych są utworzone poprzez automatyczną ich ekstrakcję z korpusu językowego, co pozwoli w przyszłości na budowę modelu automatycznie doskonalącemu się w trakcie użytkowania.

Pracę zakończono testami na zbiorze referencyjnym.

Wstęp

Przedmiotem tej rozprawy jest - zgodnie z jej tytułem - modelowanie języka naturalnego (polskiego) dla potrzeb budowy systemu rozpoznawania mowy klasy LVCSR. Rozpoczynając prezentację tej pracy warto podkreślić oryginalność jej tematu. Typowym modelem językowym stosowanymi w systemach rozpoznawania mowy dużego słownika (LVCSR), używanym głównie dla języka angielskiego jest model trigramowy [1]. Również w systemach rozpoznawania mowy tworzonych dla innych języków dominują - jak wynika z literatury - modele trigramowe i pochodne. Takie modele są w niektórych językach bardzo wygodne i wydajne,

[1] Termin ten będzie dalej dokładnie objaśniony.

ale mają tę dominującą cechę, że uwzględniają sztywno ustaloną kolejność występowania wyrazów w zdaniu. Jednak w języku polskim ta kolejność jest mniej ściśle przestrzegana niż w języku angielskim. W związku z tym przy budowie systemu klasy LVCSR, przeznaczonego do rozpoznawania mowy polskiej, potrzebny jest inny model języka, oparty na odmiennej zasadzie i dostosowany do swobodnego szyku wyrazów w zadaniu. Szyku swobodnego, charakterystycznego dla języków fleksyjnych, w szczególności dla języka polskiego. Taki właśnie model, oparty na automatycznej ekstrakcji reguł lingwistycznych, zostanie zaproponowany i przebadany w prezentowanej rozprawie. Proponowany model językowy będzie pracował zgodnie z uwarunkowaniem wynikającym ze specyficznych cech języka polskiego. Jest to jego istotna zaleta, ważna z punktu widzenia wszystkich konstruktorów systemów automatycznego rozpoznawania mowy polskiej. Drugą zaletą proponowanego modelu jest fakt, że reguły językowe zostaną w nim utworzone automatycznie poprzez ich ekstrakcję z korpusu tekstowego. Pozwoli to w przyszłości na budowę modelu, który będzie mógł podlegać automatycznemu doskonaleniu w trakcie użytkowania (będzie więc systemem uczącym się), co ze względu na zmienność języka może być uznane za istotną zaletę.

Cele i tezy pracy

Celem badań było - jak już wspomniano - opracowanie nowego modelu językowego, który mogłyby być połączony z systemem rozpoznawania mowy klasy LVCSR (z zastosowaniem dużego słownika). Model ten jest przystosowany do specyfiki języka polskiego. Oczekiwano, że zastosowanie opracowanego modelu językowego pozwoli na uzyskanie lepszych wyników rozpoznawania mowy w porównaniu z innymi modelami językowymi. Szczegółowymi celami pracy były:

1. Wybór i porównanie klas modeli językowych innych niż modele statystyczne.
2. Utworzenie prostej gramatyki powierzchniowej, będącej punktem odniesienia dla dalszych modeli.
3. Opracowanie metod oceny sekwencji słów poprzez gramatykę powierzchniową i normalizacji wyników takich ocen.
4. Implementacja symulowanego (uproszczonego) systemu rozpoznawania mowy, niezbędnego do początkowych prac.
5. Wykonanie systemu rozpoznawania mowy na bazie oprogramowania HTK.
6. Zaproponowanie sieci neuronowej mogącej uczyć się powiązań pomiędzy klasami słów, wydzielonymi na podstawie cech gramatycznych tych słów.
7. Opracowanie metod oceny sekwencji słów poprzez model sieci neuronowej i metod normalizacji takich wyników.

8. Zaproponowanie złożonej sieci neuronowej mogącej uczyć się powiązań pomiędzy klasami słów.
9. Zaproponowanie modelu sieci neuronowej mogącej uczyć się powiązań pomiędzy słowami.
10. Implementacja systemu rozpoznawania mowy pracującego na kratach słów, współpracującego z modelami autorskimi. (Modyfikacja oprogramowania HTK i SRILM)
11. Analiza testów rozpoznawania mowy.

Tezami pracy są:

1. Jest możliwe wykonanie modelu językowego dla języka polskiego z użyciem sieci neuronowych - sieci samoorganizującej się mapy Kohonena
2. Model języka naturalnego oparty na automatycznej ekstrakcji reguł pozwala uzyskać poprawę jakości rozpoznawania mowy polskiej w systemie klasy LVCSR w stosunku do szeroko stosowanego modelu bigramowego.

Zawartość pracy

Z podanych wyżej celów i tez rozprawy wynika jej struktura.

Rozdział 2 wprowadza tematykę systemów rozpoznawania mowy ciągłej z wykorzystaniem dużego zasobu słownika (LVCSR). Zostanie przedstawiona architektura typowego systemu LVCSR z uwzględnieniem roli modelu językowego. Zastosowanie modelu językowego jako element wspomnianego systemu rozpoznawania mowy wprowadza wymagania na pracę tego modelu związane ze złożonością obliczeniową a także z oceną rozpatrywanej sekwencji słów. W pracy zostaną także przedstawione wymagania stawiane takiemu modelowi z uwzględnieniem specyfiki języka polskiego, jako języka o bardziej swobodnym szyku zdania niż powszechnie badany język angielski.

Pierwszą opisywaną w pracy klasą modeli językowych będą modele statystyczne (rozdział 3). Są nimi:

- chętnie stosowany model n-gramowy (oparty na słowach) oraz jedna z jego modyfikacji
- model n-gramowy oparty na klasach.

Inne ulepszenia modelu n-gramowego zostaną tylko krótko wspomniane, bo ten model nie jest głównym przedmiotem rozważań w tej pracy. Taki wybór tematyki omawianego rozdziału pozwala na przedstawienie zasad pracy modelu n-gramowego jako modelu prostego i podstawowego dla rozpoznawania mowy. Model n-gramowy oparty na klasach słów będzie przykładem wykorzystania

związków pomiędzy klasami słów w modelu językowym, co jest istotne gdy celem jest budowa modelu opartego na klasach słów. W podrozdziale 3.3 będzie także przedstawione pojęcie miary *perplexity*, które może być użyteczne w ocenie jakości modelu językowego.

Modele n-gramowe są znane i dlatego nie będą szczegółowo dyskutowane, warto jednak je przedstawić na początku rozprawy, ponieważ będą one częściowo wzorem dla rozważanych w pracy modeli autorskich.

Kolejną klasą omawianych w pracy modeli językowych będą gramatyki formalne (rozdział 4). Będą one pozwalały na zapis powiązań pomiędzy słowami. Wprowadzenie do tematyki gramatyk formalnych obejmie przywołanie hierarchii Chomskiego opisującej pewną klasę gramatyk. Spośród nich jako modele językowe w rozpoznawaniu mowy są stosowane gramatyki liniowe. Omówiony zostanie algorytm CKY pozwalający na parsing gramatyk bezkontekstowych, gdyż w dalszych rozdziałach zostanie przedstawiona próba przeniesienia tego algorytmu na model złożonej sieci neuronowej. Przedstawiony zostanie także formalizm HPSG pozwalający na szczegółowy zapis reguł gramatycznych. Formalizm ten należy do gramatyk głębokich. Zostanie także przedstawione przeciwne do gramatyk głębokich pojęcie gramatyk powierzchniowych. Celem tego rozdziału jest wskazanie znanych rozwiązań, które będą później podlegały próbom przeniesienia na modele autorskie.

Przedstawiona w rozdziale 5 prosta gramatyka powierzchniowa jest autorskim modelem, który zawiera „ręcznie" wprowadzone najprostsze reguły gramatyczne. Mimo swojej prostoty będzie ona mogła być zastosowana jako model językowy i jej wyniki będą punktem odniesienia do modeli wymagających znacznie dłuższych przygotowań. Aby model językowy wymagał wyznaczenia jak najmniejszej liczby parametrów (wag, powiązań,...) będzie on oparty na klasach słów. To rozwiązanie przypomina model n-gramowy oparty na klasach (rozdział 3.4). Zostanie też przeanalizowana możliwość wykorzystania oceny sekwencji słów dawanej przez tą gramatykę, a następnie zostanie zaproponowany sposób normalizacji takich wyników. Inspiracją do zaproponowania takich rozwiązań był zwykły model n-gramowy oraz wspomniany wyżej model n-gramowy oparty na klasach słów. Mimo braku możliwości nauki takiego modelu, opisane rozwiązania będą częściową podbudową pod modele sieci neuronowych.

Modele oparte o sieci neuronowe (rozdział 6) są czystym opracowaniem autora niniejszej rozprawy. Zawiera ono zasady podawania informacji wejściowej dla treningu i pracy sieci, wybór sposobów kodowania wejść sieci Kohonena oraz autorskie algorytmy wykorzystania informacji z wyjść sieci Kohonena tak by uzyskać potrzebne wartości zwracane przez model językowy. Modele te są jednym z ważniejszych oryginalnych wyników naukowych zawartych w tej rozprawie. Autorskie modele będą wykorzystywać sieć samoorganizującą -mapę Kohonena do znajdowania związków pomiędzy klasami słów. Po wyjaśnieniu zasad pracy takiej sieci zostanie omówione kodowanie wejścia dla sieci. W kolejnych podrozdziałach będzie przedstawiona rozbudowa sieci, pozwalająca na znajdowanie powiązań między klasami słów i ocenę sekwencji słów. Zaproponowane będzie

postępowanie, jaki może być podjęte w przypadku niejednoznaczności klas słów, a także omówiony będzie sposób przekazania informacji o klasach słów znajdujących się w badanej sekwencji - do odpowiednich wejść sieci. Część z opisanych rozwiązań będzie zaczerpnięta z autorskiego modelu gramatyki powierzchniowej, przedstawionego w rozdziale 5. Opisany zostanie także model sieci złożonej a następnie model wykorzystujący sieć neuronową oparty o słowa. Rozdział zostanie zakończony analizą złożoności obliczeniowej i pamięciowej zaproponowanych modeli autorskich, ponieważ ta właśnie cecha stworzonych modeli może być decydująca dla ich potencjalnego przyszłego użytkownika.

Modele językowe można sprawdzić w pełni wyłącznie po podłączeniu do systemu rozpoznawania mowy. Autor takim profesjonalnym systemem nie dysponował, dlatego zachodziła potrzeba posłużenia się symulacją. Rozdział 7 opisujący zastosowany system badawczy będzie przedstawiał budowę kolejno zaimplementowanych systemów:

- systemu symulowanego rozpoznawanie mowy,
- systemu zbudowanego na bazie pakietu HTK, a także
- systemu rozpoznawania mowy wykorzystującego kraty słów.

Przedstawiony będzie także sposób znajdowania parametrów funkcji zysku, a także przytoczony będzie opis wykonania i implementacji modeli autorskich, opisanych we wcześniejszych rozdziałach. Na końcu opisane zostaną zasoby językowe wykorzystane w badaniach.

Testy modeli językowych stworzonych w ramach prac, zawarte są w rozdziale 8. Będą one przedstawiały wyniki rozpoznawania mowy uzyskane w kolejnych systemach eksperymentalnych. Najszerzej będą prezentowane eksperymenty rozpoznawania mowy wykonane w systemie wykorzystującym kraty słów (druga, ulepszona wersja krat). Działanie takiego systemu można określić jako najbardziej zbliżone do pełnego systemu rozpoznawania mowy.

Przedstawione będą także różne szczegóły dotyczące słownika, a także wyboru rozwiązań różniących modele językowe. Podany będzie opis parametrów uczenia i pracy badanych modeli. Prezentowane wyniki obejmą wykorzystanie dwóch wersji słownika różniących się wielkością. Będzie wyznaczona miara *perplexity* oraz WER (ang. procent błędnych słów) w testach rozpoznawania mowy. Obie miary będą wyznaczane w trakcie i na końcu nauki.

Rozdział 9 stanowi podsumowanie badań oraz syntetycznie przedstawia te elementy pracy, które można polecać uwadze jako oryginalny wkład autora.

Podsumowanie pracy

W pracy przedstawiono badania nad nowymi modelami językowymi, które mogą być zastosowane w systemach rozpoznawania mowy klasy LVCSR. Praca miała charakter teoretyczny (opis i analiza znanych modeli językowych, wyprowadzenie nowych modeli), jak również praktyczny (implementacja modeli językowych i systemów eksperymentalnych, trening i testy modeli).

Ważniejsze oryginalne elementy pracy są następujące: W rozdziale 6 zaproponowano modele językowe oparte o sieci neuronowe. Wyniki ich badania uzyskane w rozdziale 8 pokazują, że proponowane modele autorskie dają jakość rozpoznawania mowy (mierzoną wskaźnikiem WER) porównywalną ze znanymi modelami językowymi. Dodatkowo otrzymane wyniki pokazują, że cześć z modeli opartych o sieci neuronowe (w szczególności najlepszy z nich określony w rozdziale 8.4 jako SOMF) osiąga lepsze wyniki niż model bigramowy. Zatem spełnione są cytowane na wstępie tezy pracy.

1. Jest możliwe wykonanie modelu językowego dla języka polskiego z użyciem sieci neuronowych - sieci samoorganizującej się mapy Kohonena
2. Model języka naturalnego oparty na automatycznej ekstrakcji reguł pozwala uzyskać poprawę jakości rozpoznawania mowy polskiej w systemie klasy LVCSR w stosunku do szeroko stosowanego modelu bigramowego.

Oryginalnym wkładem autora są następujące elementy:

1. Model prostej gramatyki powierzchniowej zastosowany jako model językowy.
2. Modele klas słów oparte o sieci neuronowe samoorganizujące - mapy Kohonena , różniące się metodami kodowania wejść sieci i funkcjami wyjść sieci.
3. Model słów oparty o sieć samoorganizującą wraz z uzyskaniem funkcji ciągłej opisującej wynik zwracany przez tą sieć.
4. Model klas słów oparty o złożoną sieć samoorganizującą.
5. Zasady oceny sekwencji słów poprzez dowolne liczby wymierne dodatnie dla modeli dających ocenę zerojedynkową. Tak otrzymana miara polega także normalizacji.
6. Zasady oceny sekwencji słów w przypadku niejednoznacznego przyporządkowania słów do klas bazujących na cechach gramatycznych słów.
7. System symulowanego rozpoznawania mowy.
8. Działające implementacje wszystkich autorskich modeli językowych oraz systemów rozpoznawania mowy (pełny system oparty o HTK i system wykorzystujący kraty słów).

Wykonane (lub zmodyfikowane) w ramach pracy oprogramowanie może stanowić podstawę do dalszych badań z zakresu rozpoznawania mowy. Przedstawione wyniki prac były publikowane w czasopismach i przedstawione na konferencjach, planowane są dalsze publikacje.

Podziękowania

W sposób szczególny pragnę podziękować mojemu promotorowi prof. Ryszardowi Tadeusiewiczowi za pomoc w przeprowadzeniu badań oraz wartościowe uwagi merytoryczne w prowadzonej pracy. Wykorzystane w badaniach kraty słów zostały wykonane w Laboratorium Technologii Mowy i Języka działającego przy Poznańskim Centrum Superkomputerowo - Sieciowym. Składam tu podziękowania Pani Prof. Grażynie Demenko -kierownikowi laboratorium, a także Markowi Lange, który wygenerował wspomniane kraty.

Badania związane z tym doktoratem były finansowane w ramach Grantu promotorskiego Narodowego Centrum Nauki, nr. N516 513439, oraz poprzez stypendium programu „Podkarpacki Fundusz Stypendialny dla doktorantów". Zakup komputera na którym wykonano część badań został sfinansowany w ramach „Małopolskiego stypendium doktoranckiego".

Pragnę także podziękować dr Leszkowi Puzio z Wyższej Szkoły Informatyki i Zarządzania za udostępnienie klastra obliczeniowego. Równie serdecznie pragnę podziękować tym, którzy w mojej historii wpłynęli na moje wykształcenie oraz Rodzinie - m.in. mojej Żonie i Rodzicom za wsparcie w trakcie prac.

Summary

The subject of this dissertation is natural (Polish) language modeling for purposes of construction Large Vocabulary Continuous Speech Recognition system.

To achieve this goal theses were confirmed:

1. It is posssible to construct the language model for Polish with neural networks application - the Kohonen's Self Organized Maps.
2. The natural language model based on automatic rule extraction allows for improvement the quality of speech recognition in Polish language LVCSR systems compared to widely used bigram model.

The n-gram model typically applied in research of some languages (i.e. English) is convenient and efficient solution, but include rigid word order. Polish language needs such a class of language models that take into account free (in certain fields) word order.

Theses confirmed through the analysis of working condition of language models, the analysis of known statistical models, formal grammars, indication of the most important elements of Polish language grammars. Then shallow grammar based on simple linguistic rules was proposed. The comprehensive part of this work was devoted to author's models based on neural networks, that satisfy the conditions of free words order.

In addition, language rules in models based on neural networks are formed though automatic extraction from language corpus, which will allow in future to construction the automatically improving model during operation.

The work was completed by tests on the reference set.