Przedmowa

Niniejsza książka jest wynikiem interdyscyplinarnych (lingwistyczno-informatycznych) badań nad automatycznym dekodowaniem relacji referencyjnych w tekstach polskich. Głównych celem tych badań było stworzenie komputerowego modelu zależności tego rodzaju oraz implementacja wykrywających je narzędzi. Opisywane prace były prowadzone pod moim kierownictwem w Zespole Inżynierii Lingwistycznej Instytutu Podstaw Informatyki PAN od 2011 r. i finansowane ze środków Ministerstwa Nauki i Szkolnictwa Wyższego oraz Narodowego Centrum Nauki w ramach dwóch grantów badawczych.

Już w momencie wnioskowania o pierwszy projekt wielu kolegów, także z zagranicy, przekonywało mnie, że temat komputerowego dekodowania referencji nie jest już popularny w światowej nauce, w szczególności ze względu na spore trudności w przekroczeniu progu 70–80% miary F_1 (w zależności od języka), co w opinii niektórych możliwe byłoby tylko przy uwzględnieniu tzw. wiedzy ogólnej, wciąż trudno kodyfikowalnej w systemach komputerowych. Dodatkowy problem stanowił zamiar koncentracji prac na języku polskim, niszowym z globalnej perspektywy naukowej. Wątpliwości te potwierdziła zresztą nieudana próba nakłonienia badaczy z innych krajów do udziału w zadaniu wykrywania referencji dla polszczyzny na dostarczonych danych postawionym uczestnikom współorganizowanego przeze mnie warsztatu CORBON (*Coreference Resolution Beyond OntoNotes*) w 2016 r. Mimo wielu sygnałów wstępnego zainteresowania tematem, bariera językowa okazała się zbyt wysoka lub wyniki uzyskiwane standardowymi metodami zbyt słabe, by je zaprezentować.

Dekoder zależności referencyjnych stanowił jednak ważny element, którego brakowało w zestawie podstawowych narzędzi językowych powstałych w ostatnich latach dla polszczyzny. Mogłyby z niego w oczywisty sposób skorzystać algorytmy automatycznego streszczania (np. w celu zastępowania wyrażeń niepełnoznacznych), tłumaczenia komputerowego (do ujednoznaczniania wariantów tłumaczeń) czy analizy metatekstowej. Jednocześnie w ciągu ostatnich lat nastąpił intensywny rozwój nowych, efektywnych metod komputerowych, a zaspokojenie „pierwszych potrzeb" w dziedzinie polskiej inżynierii lingwistycznej umożliwiło skoncentrowa-

nie prac na bardziej wymagających problemach z pogranicza składni i semantyki oraz referencji oraz dyskursu (metatekstu).

W związku z tym, że w języku polskim zagadnienie przetwarzania relacji referencyjnych w ujęciu ogólnym nie było dotąd systematycznie badane metodami lingwistyczno-informatycznymi, praca ta stanowi pierwszą skondensowaną próbę komputerowego opisu referencji nominalnej w języku polskim oraz przedstawienie sposobu implementacji narzędzi do jej wykrywania. Zgodnie z aktualnymi trendami wykorzystuję do tego celu podejście korpusowe, z ręczną anotacją konstrukcji referencyjnych, pozwalające zarówno na weryfikację zaproponowanej teorii na rzeczywistych danych, jak i tworzenie narzędzi automatycznych metodami maszynowego uczenia, a następnie ocenę jakości powstałych narzędzi za pomocą standardowych miar ewaluacyjnych.

Książka podzielona jest na części odpowiadające głównym blokom tematycznym pracy korpusowo-informatycznej. Po przedstawieniu założeń (rozdział 1) oraz stanu obecnej wiedzy teoretycznej i praktycznej w zakresie, w jakim była przydatna w pracach algorytmicznych (rozdział 2), prezentuję stworzony na ich potrzeby model relacji referencyjnych (rozdział 3), użyty następnie w procesie anotacyjnym o szczegółowo określonych ramach, który doprowadził do powstania korpusu zależności referencyjnych (rozdział 4). Dane korpusu posłużyły następnie do stworzenia kilku wariantów narzędzi do automatycznego wykrywania referencji (rozdział 5), a ich jakość została oceniona zgodnie z dostępnymi metrykami (rozdział 6). Perspektywa dalszych badań (rozdział 7) została zaprezentowana w szerszym kontekście modelowania relacji metatekstowych. Ostatni rozdział stanowi krótkie podsumowanie uzyskanych wyników.

Obecna publikacja prezentuje czytelnikowi polskiemu prace prowadzone w trakcie ośmiu lat, co wiąże się z dwiema konsekwencjami. Pierwszą z nich jest konieczność podsumowania wyników opisywanych już częściowo wcześniej, w monografii anglojęzycznej (Ogrodniczuk 2015) oraz licznych artykułach i publikacjach konferencyjnych. Drugą – potrzeba skondensowanego przedstawienia obszernego materiału. W celu ułatwienia lektury wszystkie fragmenty, mogące wymagać dokładniejszych objaśnień, zostały zaopatrzone w odesłania do wcześniejszych prac. Na końcu książki zamieszczono jej angielskie streszczenie przeznaczone dla czytelników zagranicznych.