Ramy semantyczne w tekstach prawnych – analiza konfrontatywna polsko-angielsko-niemiecka

Agnieszka Pluwak

Cele badawcze pracy

Analiza leksykalna słownictwa związanego z pojęciem *najmu*, niezbadanego dotąd metodą ramową, a zatem:

1. wkład w budowę FrameNetu dla języka polskiego; pośrednio – przyczynienie się do rozwoju zasobów leksykalnych, stosowanych w przetwarzaniu języka naturalnego;

2. przygotowanie materiału językoznawczego (leksyka, elementy ramy, konstrukcje) do zakodowania, czyli przełożenia na język reguł informatycznych, celem stworzenia programu do automatycznego wyszukiwania informacji w tekstach umów prawnych.

Tezy:

Główne tezy pracy są ujęte w jej poszczególnych rozdziałach:

1. *Teorię semantyki ramowej w ujęciu Fillmore'a da się zastosować do analizy wielu języków, gdyż postulowane przez niego ramy mają charakter pojęciowy (semantyczny).*
2. *Zastosowanie semantyki ramowej jako metody do analizy języka polskiego:*
   A. rodzi dylematy natury formalnej - w dopasowaniu do istniejących już ram ze względu na cechy charakterystyczne dla budowy języka słowiańskiego takie, jak aspekt czasownika, rozbudowane prefiksy, podmiot domyślny zawarty w czasowniku - ale nie rodzi problemów natury pojęciowej ze względu na ogólny charakter definicyjny ram;
   B. stwarza konieczność zbudowania nowych ram dla niektórych nieopisanych dotychczas we FrameNecie znaczeń;
   C. daje możliwość zastosowania wyników analizy w tworzeniu systemów NLP[1].
3. *Semantyka ramowa w ujęciu FrameNet znajduje zastosowanie w wielu funkcjach przetwarzania języka naturalnego:*
   A. jako projekt leksykalny (oznaczony przez lingwistów korpus tekstów prasowych służący do trenowania systemów uczenia maszynowego),
   B. poszczególne opisane ramy znajdują zastosowanie w funkcjach ekstrakcji danych typu oznaczanie ról semantycznych (ang. semantic role labeling) jako schematy będące szkieletem całego systemu NLP.
4. *W badaniach językoznawczych (zarówno językoznawstwa tradycyjnego jak i komputerowego) istniało wiele podejść do analizy argumentów (walencja syntaktyczna i semantyczna) i jej zastosowania. Jednak w przypadku analiz dziedzinowych takich, jak język prawniczy*

[1] NLP (ang. Natural Language Processing) – przetwarzanie języka naturalnego, zwane też językoznawstwem komputerowym.

opisujący przypadek najmu, potrzebny jest opis walencji semantycznej (a nie jedynie syntaktycznej) na bardziej szczegółowym poziomie (niż VerbNetu czy PropBanku) i dlatego wybrano podejście semantyki ramowej w ujęciu FrameNet do analizy wybranego materiału.

Powyższe tezy te składają się na tezę główną, która brzmi:

Możliwym jest wpisanie leksemów z różnych języków, w tym – dla języka polskiego - w ramy semantyczne zdefiniowane w projekcie FrameNet lub zbudowanie nowych ram semantycznych dla słownictwa prawniczego, pojawiającego się w umowach najmu, na ich podstawie – zdefiniowanie konstrukcji semantyczno-gramatycznych, którymi wyrażane są poszczególne elementy ramy, a następnie przełożenie ich na reguły systemu informatycznego.

Materiały badawcze

Do analizy wybrano słownictwo prawnicze (około 20-30 jednostek leksykalnych, głównie czasowników, ale też rzeczowników, przymiotników i przysłówków) z dziedziny najmu nieruchomości. O doborze materiałów zadecydowały następujące czynniki:

A. fakt, iż słownictwo to nie zostało dotychczas opisane w projekcie RAMKI, co umożliwia poszerzenie zasobów projektu, mającego szansę stać się początkiem polskiego FrameNetu;

B. dostępność materiałów w ww. językach, to znaczy prawnych umów najmu,

C. możliwość praktycznego wykorzystania projektu do automatyzacji analizy umów najmu w jednej z największych firm z branży zarządzania nieruchomościami w Europie.

Metody badawcze

Semantyka ramowa w ujęciu FrameNet jest podejściem szeroko stosowanym w ekstrakcji danych z tekstów, gdyż jest to projekt leksykograficzny, który od podstaw zakładał wykorzystanie swoich zasobów w lingwistyce komputerowej. Analiza w ujęciu FrameNet tym różni się od innych podejść do analizy argumentów, że:

1. jest to analiza głównie semantyczna na poziomie szczegółowym (jest to opis na poziomie kategorii kognitywnych, w którym przypisuje się elementom ramy konkretne role sytuacyjne, a nie na bardziej ogólnym poziomie agenta, jak ma to miejsce w innych projektach leksykalnych np. VerbNet);

2. jest to analiza również syntaktyczna, która pokazuje, jak elementy składniowe realizują poszczególne role semantyczne;

3. jest to zapis umożliwiający zebranie jednostek leksykalnych w zbiory o wspólnej części znaczenia.

Metodę ramową wybrano do analizy materiału, gdyż realizacja celów pracy wymaga podejścia bardziej szczegółowego niż w innych projektach leksykograficznych, a podejście FrameNet zawiera zarówno semantykę, jak i konstrukcje syntaktyczne, umożliwiające przekład na system reguł do przetwarzania informatycznego.

Konstrukcja pracy:

Część teoretyczna

Rozdział 1 (około 10 - 20 stron)

O pojęciu ramy (ang. *frame*) w różnych dyscyplinach naukowych. Ramy w lingwistyce (Fillmore oraz Lakoff) i kontrowersyjne pytanie o to, czy ramy Fillmore'a są konceptualne, czy tylko leksykalne i jeśli tak, to dlaczego i jakie ma to implikacje.

Rozdział 2 (około 20 stron)

Rozwój teorii semantyki ramowej Fillmore'a – od rozróżniania znaczenia dla takich samych składniowych konstrukcji walencyjnych (w opozycji do gramatyki generatywnej Chomsky'ego), aż po różne podejścia do badań ról semantycznych w lingwistyce. Zdefiniowanie popularnego zestawu ról semantycznych w dziedzinie informatyki, badania np. Langackera nad rozwojem przypisywania ról semantycznych (ang. semantic role labeling), aż po rozwinięcie teorii semantyki ramowej Fillmore'a do kształtu projektu FrameNet, czyli jej zastosowania w funkcjach przetwarzania języka naturalnego.

Rozdział 3 (około 40 stron)

Dotychczasowe osiągnięcia i dylematy FrameNetów dla języka polskiego, angielskiego i niemieckiego: wyniki (ile zostało opisanych jednostek leksykalnych, ile powstało ram, czy trzeba było tworzyć nowe ramy i jeśli tak, to dlaczego? Jakie były różnice w wynikach badań w poszczególnych językach i z czego one wynikały?)

Rozdział 4 (około 30 stron)

Możliwości zastosowania semantyki ramowej w ujęciu FrameNet w funkcjach przetwarzania języka naturalnego takich, jak budowa dziedzinowych systemów reprezentacji wiedzy, automatyczne wyszukiwanie danych w tekstach, czy też automatyczna analiza semantyczna z zastosowaniem kategorii pojęciowych.

Analiza

Rozdział 5 (około 30 stron)

Metodologia opisu ram w projekcie FrameNet: jak wygląda system oraz proces anotacji, jakie są jego zasady. Krytyka systemu opisu ram we FrameNecie.

Rozdział 6 (około 100-150 stron)

Wstępnie do analizy wybrano następujące jednostki leksykalne:

Czasowniki: mieszkać, zawrzeć (umowę), zwać (nazywać), zobowiązać się, płacić / wypłacić, wyrazić (zgodę) , móc, wypowiedzieć, zwrócić, dokonać, sprzedawać, etc.

Rzeczowniki: umowa, najemca, wynajmujący, wy/najem, podnajem, obowiązek, odpowiedzialność, etc.

Przyimki: od... do..., w (dniu), pomiędzy, etc.

Są to tylko przykłady jednostek leksykalnych, których pełen zestaw można będzie podać po zakończeniu procesu analizy. Podobnie jak w projekcie RAMKI, na opis ramowy składały będą się: walencja morfo-syntaktyczna i syntaktyczna (typowe części mowy, zdania i konstrukcje wyrażające dane semantyczne argumenty czasownika, rzeczownika, przymiotnika lub przysłówka), definicja pojęcia dla ramy (schematyczny opis sytuacji), jego ról obowiązkowych i dowolnych (walencja semantyczna), przykłady wyrażających daną ramę jednostek leksykalnych z wybranych umów, oznakowane na przykładach zdań z Narodowego Korpusu Języka Polskiego.

Rozdział 7 (około 30 stron)

Wyniki przeprowadzonej analizy – zestawienie kontrastywne ram i jednostek leksykalnych, różnice i podobieństwa w konstrukcjach gramatycznych i leksemach wyrażających poszczególne elementy ramy w danym języku.

Zbudowanie całościowego schematu relacji między ramami.

Rozdział 8 (około 30 stron)

Zastosowanie wyników analizy do zbudowania aplikacji do ekstrakcji danych z umów najmu w wybranym języku. Pokazanie zastosowania analizy ramowej w funkcjach NLP. Zbudowanie reguł informatycznych na bazie konstrukcji gramatycznych zawartych w ramach.

Wnioski

Rozdział 9 (około 20 stron)

Potwierdzenie ww. tez wraz z wyszczególnieniem przykładów modyfikacji lub budowy nowych ram semantycznych dla badanej dziedziny. Wyszczególnienie konstrukcji gramatycznych języka polskiego, różniących się od konstrukcji dla pozostałych badanych języków. Opis występujących podobieństw i różnic pojęciowych. Porównanie budowy powstałej ontologii dziedzinowej z projektem ontologii medycznej powstałej w oparciu o semantykę ramową. Opis konstrukcji gramatycznych do przepisania w formacie reguł informatycznych.

Znaczenie spodziewanych wyników badania dla rozwoju językoznawstwa

1. powiększenie zasobów jednostek leksykalnych języka polskiego, zanalizowanych metodą ramową (projekt FrameNet jest obecnie jednym z najważniejszych projektów leksykalnych językoznawstwa komputerowego na świecie i brak zasobów dla języka polskiego oznacza brak możliwości uczestnictwa w niektórych projektach budowy światowych systemów translacji maszynowej);
2. wkład w rozwój badań kontrastywnych metodą ramową (o zastosowaniu w teorii translacji, konfrontacji językoznawczej i lingwistyce korpusowej);

3. zbudowanie lingwistycznej podstawy systemu informatycznego do ekstrakcji danych z tekstów umów prawnych, opartego o funkcję anotacji ról semantycznych (ang. *semantic role labelling*) na bazie zdefiniowanych uprzednio ram semantycznych.