Konrad Gołuchowski
Metoda identyfikacji schematów walencyjnych w tekstach polskich
Rozprawa doktorska

Streszczenie

Rozprawa porusza temat automatycznej identyfikacji schematów walencyjnych w tekstach polskich. Schematy walencyjne są to struktury, które określają, w jaki sposób argumenty predykatu mogą być realizowane w zdaniach. Wyniki identyfikacji schematów walencyjnych mogą zostać wykorzystane m.in. w zadaniu wykrywania zdarzeń (ang. *event detection*), które ma liczne zastosowania praktyczne. Interesującym polem badań naukowych jest także wykorzystanie informacji o występującym w zdaniu schemacie walencyjnym do poprawy skuteczności parsowania składniowego.

Celem opisanych badań jest zbudowanie systemu do automatycznej identyfikacji schematów walencyjnych bez wykorzystania ręcznie przygotowanego korpusu treningowego oznaczonego tego rodzaju schematami. Takie podejście do identyfikacji schematów walencyjnych jest istotne, gdyż przygotowanie korpusu treningowego obejmującego czasowniki powszechnie wykorzystywane w języku codziennym, który nadawałby się jako korpus treningowy, jest w zasadzie niewykonalne w rozsądnym czasie ze względu na mnogość czasowników i ich schematów walencyjnych. Warto zauważyć, że jest to pierwsza próba identyfikacji schematów walencyjnych dla języka polskiego i w zasadzie jedyna – dla jakiegokolwiek języka – zaawansowana próba identyfikacji schematów walencyjnych bez wykorzystania ręcznie oznaczonego korpusu treningowego.

Cel pracy został osiągnięty poprzez opracowanie nowej metody automatycznej identyfikacji predykatów i ich podrzędników, automatyczną budowę zbioru przykładów użyć schematów walencyjnych oraz opracowanie siedmiu algorytmów wyboru schematów walencyjnych. Automatyczne zbudowanie zbioru przykładów umożliwia obliczenie preferencji selekcyjnych, czyli informacji semantycznych opisujących sensowne realizacje argumentów. Ponadto zbiór ten służy jako korpus treningowy do budowy algorytmów uczących się identyfikacji schematów walencyjnych, wyko-

rzystujących automatycznie wydobyte preferencje selekcyjne, wektorową reprezentację słów oraz informację o podrzędnikach predykatów. Do wyboru schematów walencyjnych zastosowano zarówno klasyczne klasyfikatory, m.in. naiwny klasyfikator Bayesa, regresję logistyczną, lasy losowe, jak i autorskie algorytmy opracowane specjalnie do wyboru schematów walencyjnych. Wszystkie te algorytmy zostały złożone w jedną metodę za pomocą metaheurystyki symulowanego wyżarzania.

Dodatkowym wkładem w rozwój narzędzi przetwarzania języka naturalnego jest opracowana i zaimplementowana metoda identyfikacji podrzędników predykatów. Może ona zostać zastosowana w wielu zadaniach przetwarzania naturalnego, m.in. oznaczania argumentów rolami semantycznymi, wydobywania informacji, automatycznego tworzenia słowników walencyjnych. Zarówno kod źródłowy opracowanych algorytmów, jak i wytrenowane modele, zostały opublikowane i są dostępne w internecie. Szczegóły ich wykorzystania zostały opisane w dodatku D.

W rozprawie opisano także szereg przeprowadzonych eksperymentów oceniających skuteczność zaproponowanych rozwiązań. Uzyskane wyniki są znacznie lepsze niż rezultaty jedynej wcześniejszej, bardzo prostej, metody identyfikacji schematów walencyjnych bez nadzoru przeprowadzonej dla języka czeskiego i zbliżają się do najlepszych wyników uzyskiwanych dla języka czeskiego przez algorytmy uczące się z nadzorem.

Konrad Gołuchowski
Metoda identyfikacji schematów walencyjnych w tekstach polskich
Rozprawa doktorska

Abstract

The dissertation deals with the issue of automatic identification of valence patterns in Polish texts. Valence schemes are structures that define how predicate arguments can be realized in sentences. The results of the identification of the validation schemes may be used, for example, in the event detection task, which has many practical applications. An interesting field for scientific research is also the use of information on the used valence scheme in a sentence to improve the effectiveness of syntax parsing.

The goal of the study is to build a system for automatic identification of valence patterns without the use of a manually prepared training body marked with such patterns. This approach to the identification of valence patterns is important as it is not feasible to prepare a training corpus containing verbs commonly used in everyday language, which would be suitable as a training corpus, within a reasonable time due to the multiplicity of verbs and their valence patterns. It is worth mentioning that this is the first attempt to identify valence schemes for the Polish language and basically the only – for any language – advanced attempt to identify valence schemes without the use of a manually annotated training corpus.

The aim of the research was achieved by developing a new method of automatic identification of predicates and their dependents, automatic construction of a set of uses of valence schemata and development of seven algorithms for selection of valence schemes. Automatic construction of a set of examples enables calculation of selection preferences, i.e. semantic information describing sensible realization of arguments. In addition, this set serves as a training corpus for the construction of algorithms learning to identify valence schemes, using automatically extracted selection preferences, vector representation of words and information about predicate dependents. Classic classifiers, such as the naïve Bayes classifier, logistic regression, random forests, as well as algorithms developed specifically for the selection of valence schemes, were used to select valence schemes.

All these algorithms have been combined into a single method using the simulated annealing metaheuristic.

The method of identification of predicate subordinates has been developed and implemented as an additional contribution to the development of natural language processing tools. It can be used in many natural processing tasks, such as semantic roles labelling, information extraction, automatic creation of valence dictionaries. Both the source code of the developed algorithms and the trained models have been published and are available on the Internet. Details of their use are described in Appendix D.

A number of experiments evaluating the effectiveness of the proposed solutions are also described in the dissertation. The obtained results are much better than the results of the hitherto only, very simple, unsupervised method of identification of valence schemes for Czech, and are approaching the best results for the language Czech achieved by supervised learning algorithms.