Streszczenie

Niniejsza praca dotyczy znanego w dziedzinie przetwarzania języka naturalnego problemu odpowiadania na pytania (ang. *Question Answering*, QA). Zadanie to polega na budowie w pełni automatycznego systemu komputerowego zdolnego do przyjmowania pytań w języku naturalnym (np. polskim) i udzielania odpowiedzi w tym samym języku. Problem ten pełni szczególnie istotną rolę w przypadku systemów komputerowych udostępniających informacje osobom o niskich kwalifikacjach informatycznych (np. w przypadku wyszukiwarek internetowych).

W pracy przedstawiono opis systemu QA dla języka polskiego, nazwanego RAFAEL (ang. *RApid Factoidal Answer Extracting aLgorithm*), zdolnego do interpretacji pytań o proste fakty, odnajdywania informacji w zbiorze dokumentów tekstowych bez ograniczeń dziedzinowych i udzielania zwięzłych odpowiedzi. Pewne własności języka polskiego, tj. fleksyjność i swobodny szyk zdania, sprawiają, że niektóre elementy zadania (np. wstępny wybór dokumentów i dopasowywanie zdań) wymagają specjalnego podejścia i sprawiają trudności nieobecne w przypadku języka angielskiego, dla którego budowana jest większość systemów QA.

W odróżnieniu od innych badań z tej dziedziny, w niniejszej pracy najwięcej uwagi poświęcono problemowi rozpoznawania nazw. Zamiast powszechnie stosowanego podejścia, bazującego na rozpoznawaniu nazw własnych (ang. *Named Entity Recognition*, NER), zaproponowano jego uogólnienie, nazwane głębokim rozpoznawaniem nazw (ang. *Deep Entity Recognition*, DeepER). W tej wersji odnajdywanym w tekstach nazwom, potencjalnie stanowiącym odpowiedzi, nie są przypisywane szerokie kategorie NE, ale szczegółowe synsety z ontologii WordNet. Pozwala to na precyzyjniejsze wybieranie kandydujących wzmianek i zapewnienie pełnej zgodności z ograniczeniami wyrażonymi w pytaniu. Co więcej, poszerza to zakres obsługiwanych pytań o te dotyczące bytów takich jak zwierzęta, urządzenia czy związki chemiczne, leżących poza tradycyjnymi kategoriami NE.

Ewaluacja systemu przeprowadzona została w dwóch krokach – w pierwszym wykorzystano zbiór 1130 pytań z teleturnieju *Jeden z dziesięciu*, podczas gdy treść polskiej Wikipedii posłużyła za korpus tekstów. Dzięki technice DeepER, możliwe stało się także automatyczne sprawdzenie poprawności tych odpowiedzi, które sformułowano inaczej niż wzorcowe. Pozwoliło to na przeprowadzenie serii eksperymentów mierzących poprawność odpowiedzi w różnych konfiguracjach. W drugiej części uzyskane optymalne parametry wykorzystano w ręcznej ewaluacji wydajności systemu RAFAEL na oddzielnym zbiorze 576 pytań. Uzyskane wyniki pokazują, że użycie zaproponowanego podejścia znacząco poprawia udzielane odpowiedzi, głównie poprzez uwzględnienie pytań niedostępnych dla dotychczasowych metod.

Abstract

Title: *Question answering in Polish using deep entity recognition*

This thesis addresses a well-known problem in the domain of Natural Language Processing (NLP): Question Answering (QA). The task is to create a fully automatic computer system, capable of accepting questions in natural language (e.g. Polish) and returning answers using the same language. The problem plays an important role in computer systems that are expected to be used by people of low computer skills (e.g. search engines).

The work presents a description of a QA system for Polish, called RAFAEL (RApid Factoidal Answer Extracting aLgorithm), which can interpret factoid questions, find answers in a plain-text open-domain corpus and formulate them in concise and short manner. Some of the properties of Polish, i.e. rich nominal inflection and free word order make NLP sub-tasks, e.g. relevant documents selection and sentence matching, more challenging.

Unlike in other studies in the domain, the focus of this work is on a problem of entity recognition. In place of traditional Named Entity Recognition (NER) approach, its generalisation, called Deep Entity Recognition (DeepER), is proposed. Instead of using few general NE categories, entities are described by a set of WordNet synsets, to which they belong. It leads to much more precise candidate mention selection, according to question constraints. Moreover, a range of answered questions broadens, as those beyond the traditional NE categories (e.g. animals, devices, chemical compounds) are also accepted.

System evaluation consists of two steps – in the first part a set of 1130 questions from a Polish TV quiz show is used, while contents of Polish Wikipedia serve as a knowledge base. Thanks to DeepER technique, it is possible to automatically check validity of an answer even when it is formulated differently than expected. The method let to perform a series of experiments, measuring answering accuracy in varying configurations. Secondly, the obtained parameters were used to manually evaluate RAFAEL performance on previously unseen 576 questions. The results show that using DeepER approach substantially improves the answers, mainly by handling questions lying beyond the reach of existing methods.