Mateusz Kopeć

Summarization of Polish Press Articles Using Coreference

PhD dissertation

Supervisor: dr hab. Agnieszka Mykowiecka
Auxiliary supervisor: dr Maciej Ogrodniczuk

Institute of Computer Science
Polish Academy of Sciences
Warsaw, 2018

Streszczenie

W dzisiejszych czasach ilość dostępnych informacji w formie tekstowej jest ogromna, głównie ze wzgledu na ich elektroniczną formę przechowywania. Trudno jest nadążyć z przetwarzaniem nowopowstających źródeł tekstowych, produkowanych w bardzo dużym tempie. Informacje tekstowe są czesto podane w bardzo rozbudowanej postaci. Automatyczne streszczanie pozwala na pominięcie nadmiarowych informacji, w momencie gdy zależy nam tylko na kluczowych faktach. Jest to proces wyboru i transformacji dostępnych danych w celu zaprezentowania ich w bardziej treściwej formie, zachowując jednak najistotniejsze fragmenty oryginalnego źródła.

Dotychczas automatyczne streszczanie tekstów w języku polskim doczekało się niewielkiej liczby badań. Celem niniejszej pracy doktorskiej była porównawcza analiza istniejących podejść do problemu automatycznego streszczania i zaproponowanie nowej metody. Kluczową cechą opracowanej metody jest opieranie estymacji istotności fragmentów tekstu źródłowego głównie na informacji o koreferencji w tekście. Dane o powiązaniach między wskazywanymi w zdaniu obiektami wspomagają proces wyboru najistotniejszych fragmentów tekstu, wybieranych do tworzonego automatycznie streszczenia. Niezbędne narzędzia i dane dotyczące wykrywania koreferencji w tekstach w języku polskim pojawiły się w ostatnich latach w związku z intensywnymi pracami kilku grup badawczych nad tym tematem.

Tezę badawczą pracy stanowiło stwierdzenie, że automatycznie wykryta koreferencja pozwoli poprawić wyniki automatycznego systemu streszczającego polskie artykuły prasowe, tak aby działał ze skutecznością przekraczającą najlepsze aktualnie dostępne rozwiązania. Główny naukowy wkład niniejszej pracy stanowi opracowanie i implementacja najbardziej skutecznego algorytmu streszczania polskich tekstów prasowych, opierającego się w głównej mierze na wykorzystaniu informacji o koreferencji w tekście źródłowym. Po przeprowadzeniu licznych eksperymentów, opracowana została trójetapowa procedura umożliwiająca efektywne rozwiązanie problemu możliwych błędów powstających przy wykorzystaniu automatycznie wydobytych relacji koreferencji. Pierwszy jej etap to klasyfikacja fraz będących odniesieniami

do obiektów pod kątem ich istotności dla streszczenia. Drugi etap to ocena ważności zdania z uwzględnieniem relacji koreferencji. Trzeci krok to wybór najważniejszych zdań do stworzenia ekstraktu o zadanej długości. W celu przeprowadzania dwóch pierwszych etapów wykorzystywane są modele powstałe przez uczenie maszynowe. Zaproponowana została nowa metoda trenowania takich modeli, wykorzystująca tzw. "optymalne streszczenie", czyli sztucznie stworzoną syntezę wielu ręcznie wyprodukowanych streszczeń danego tekstu.

W prezentowanej pracy opisano też wyniki badań dotyczące automatycznego poprawiania generowanych streszczeń: pojawia się w niej pierwsze podejście do automatycznego tworzenia podmiotów zerowych w tekstach w języku polskim. Zbadano możliwość usuwania podmiotu w zdaniu w wypadku, gdy w zdaniu poprzedzającym pojawia się ten sam podmiot, a jego usunięcie z kolejnego zdania nie spowoduje zmiany znaczenia.

Praca zawiera szeroki przegląd prób wykorzystywania koreferencji w automatycznym streszczaniu, jak również porównanie wyników wszystkich dostępnych publicznie systemów streszczających polskie teksty. Ponadto, dla języka polskiego zostały zaadaptowane dwa wiodące systemy dla języka angielskiego: oparte na podejściu grafowowym i sieciach neuronowych. Co więcej, dla jednej z metryk ewaluacyjnych obliczony został najwyższy teoretycznie możliwy do uzyskania wynik. Został on porównany do wyników osiągniętych przez ludzi, streszczających źródłowe teksty ręcznie. Pozwoliło to na szersze zrozumienie znaczenia rezulatów aktualnie uzyskiwanych przez automatyczne systemy streszczające. Zaproponowana została również nowa miara – *ROUGE-M*, nie posiadająca jednej ze słabości pozostałych miar z rodziny *ROUGE*. *ROUGE-M* zakłada, że dla danego tekstu może istnieć wiele dobrych streszczeń, używających odmiennych słów do podsumowania treści.

Na potrzeby ewaluacji stworzonych algorytmów utworzony został POLSKI KORPUS STRESZCZEŃ 'POLISH SUMMARIES CORPUS'. Jest to największy obecnie korpus streszczeń zawierający teksty w języku polskim, stanowiący dużą wartość dla przyszłych badań w temacie automatycznego streszczania. Autor niniejszej pracy włożył wiele wysiłku w projekt procedury anotacyjnej, koordynację ręcznych prac nad korpusem i ostatecznie publikację zasobu dla wszystkich zainteresowanych.

Tytuł pracy w języku polskim

Streszczanie polskich artykułów prasowych z wykorzystaniem koreferencji

Abstract

Nowadays, the amount of available text information is astonishingly high, mainly due to its electronic form of storage. There is no possibility to read and monitor all the new data, which is produced at a very high pace. However, available text information is often not short and concise. Automatic summarization allows to omit the excessive content, when we want to focus only on the most important facts. It is the process of selection and transformation of available data with the aim of presenting it in a more succinct way, at the same time conveying the most important fragments of the original source.

As until now, not much research was done in the area of automatic summarization for Polish language texts. The motivation for this thesis was to perform a comparative analysis of existing approaches of summarization and propose a new method. The key feature of the new method was to rely mainly on coreference information for estimating the importance of text fragments, which then could be composed to form an automatic summary. Automatic coreference resolution for the Polish language has recently been studied by several research groups. The mentioned research resulted in providing necessary data and tools.

The research thesis of this work based on the notion that automatic coreference resolution may be used to create an automatic summarizer for Polish press articles, outperforming current state-of-the-art. The main contributions of the thesis is the design and implementation of the currently best Polish news summarization algorithm, which heavily relies on using coreference information. After many experiments, a three-step procedure was developed to efficiently cope with a non-perfect coreference information output from automatic coreference resolvers. Its first stage consists of classification of object-referring phrases regarding their importance for the summarization process. The second stage is to weigh each sentence from the input text according to its salience, using the coreference information. The final step is to compose an extractive summary from the sentences with the highest scores. The first two stages use machine-learned models. For training these models, a custom training

method was invented, relying on a so-called 'optimal summary', an artificially created synthesis of many manually crafted summaries.

The automatic summary revision field has also been extended by this thesis: we showed the first attempt in the literature to automatically introduce zero subjects in Polish language texts. We proposed and evaluated the possibility to remove the subject from a sentence in case the preceding sentence contains the same subject. Of course, after the removal the meaning of the text should be left intact.

An extensive survey of automatic summarization systems related to coreference is presented in the thesis, together with the evaluation of all publicly available summarization systems for Polish texts. Two new systems for the Polish language were created by adapting state-of-the-art English algorithms: a graph-based and a neural network-based one. Moreover, for one of the evaluation measures, we calculated the theoretically obtainable upper bounds for each text in the corpus and, at the same time, compared it with scores of human annotators, giving a 'bigger picture' of what current evaluation scores mean. Additionally, we proposed a new metric: *ROUGE-M*, which overcomes one of the weaknesses of the original *ROUGE* metric family. *ROUGE-M* implements the assumption, that a single text may be summarized in various ways, each one using different vocabulary.

To perform a meaningful evaluation, POLISH SUMMARIES CORPUS was developed. It is the largest corpus with summaries in the Polish language and a resource of great value for future evaluations of automatic summarization systems. The design of the annotation procedure, coordination of manual annotations and the final release of the corpus to the public were all steps which required a large effort from the author of this thesis.

Thesis title

Summarization of Polish Press Articles Using Coreference