MAGDALENA ZAWISŁAWSKA

O problemach z koreferencją

Celem niniejszego artykułu jest omówienie niektórych problemów, które występują podczas określania koreferencji w tekście. Analizy takie były prowadzone na potrzeby projektu CORE – Komputerowe metody identyfikacji nawiązań w tekstach polskich (kierowanego przez Macieja Ogrodniczuka). Głównym celem projektu było stworzenie nowatorskich metod i narzędzi informatycznych służących do automatycznego wykrywania anafor i koreferencji w tekstach pisanych w języku polskim.
Główny problem z wyznaczaniem faz koreferencji w języku polskim wyłonił się w efekcie kilku czynników. Na poziomie pragmatycznym i semantycznym nie było proste zadecydować czy zachodziła identyczność, czy tylko podobieństwem między dwoma obiektami. Dodatkowym utrudnieniem był brak specjalistycznej wiedzy, który sprawił, że wyznaczenie faz koreferencji było szczególnie trudne między frazami w wyjątkowo specjalistycznych tekstach. Na poziomie gramatycznym, niektóre cechy języka polskiego utrudniły anotację. Ze względu na brak rodzajników określonych i nieokreślonych bardzo trudno było określić, czy nadawca zawsze miał na myśli ten sam obiekt, czy różne obiekty należące do tej samej klasy. Wreszcie, długie zdania bez podmiotu spowodowały pewne problemy przy wyznaczaniu łańcuchów koreferencyjnych między analizowanymi frazami.

Słowa kluczowe:
referencja, koreferencja, korpus, anotacja

About problems with coreference

The aim of the paper is the presentation of some problems occurring during coreference annotation. Such an analysis was performed for the project *CORE – Computer-based methods for coreference resolution in Polish texts* (managed by Maciej Ogrodniczuk). The project's main goal was to create innovative methods and tools for automated anaphora and coreference resolution in Polish texts.
The main problem with the coreference resolution in the Polish language arose due to several reasons. On the pragmatic and semantic level it wasn't easy to decide if there was an identity or just a similarity between two different objects. Another problem was the lack of specific knowledge which made it a very hard task for the annotator to see the coreference between phrases in some highly specialist texts. On the grammatical level, some properties of the Polish language made the annotation difficult. There are no definite and indefinite articles in Polish, therefore it was very hard to determine if the speaker had meant always the same object or just different specimens belonging to the same class. Also, long subject-less sentences presented some problems with defining the coreference chains between analyzed phrases.

Keywords:
reference, coreference, corpus, annotation

Magdalena Zawisławska
Uniwersytet Warszawski
zawisla@uw.edu.pl