

Multilingual Events Extraction

Abstract

Many events are instantly spread online on the internet from all over the world. In most of the times these events emerged as a fraction of multi event hierarchy or just standalone events. For example, in the last four years a main event started in Arabic region, which is known as "The Arab Spring". Since that event started everyone tried to follow and catch any development. The main goal for this thesis is to present a new application for automatic multilingual event extraction from text (OraMEE), such an even being, for example, the above mentioned one.

One main task and meanwhile application for OraMEE is detecting any conflict-related event from a multilingual corpus called EventCorp and then storing these events in a database tables (OraMEE schema). Furthermore, OraMEE approach is a novel technique for extracting a specific event based on a predefined domain "Arabs Spring". As any Information Extraction (IE) system, OraMEE needs a suitable annotated corpus to train and evaluate each OraMEE components and sub-process. Furthermore, the selected corpus must mainly contain events.

Due to this fact, it was very difficult to find a proper corpus that contains Arab Spring daily events, there was a need to develop and build such a corpus. Therefore and for the purposes of achieving our target, which is a OraMEE, we collected some of those events in what we called the EventCorp corpus and we presented the results in a paper (Alasfour & Trausan-Matu, 2013). Furthermore, it was needed for EventCorp to be aligned (English, Arabic) and annotated (with Named Entities), therefore, a 3D aligned approach was developed for the purpose of EventCorp Alignment (Alasfour & Trausan-Matu, 2014).

Extracția Evenimentelor Multilingve

Rezumat

Multe evenimente sunt răspândite instantaneu on-line pe internet din întreaga lume. În cele mai multe ori aceste evenimente au apărut ca o fracțiune de multe ierarhie eveniment sau evenimente doar de sine stătătoare. De exemplu, în ultimii patru ani un eveniment principal a început în regiunea arabă, care este cunoscut sub numele de "Primăvara arabă". De la început acest eveniment toată lumea a încercat să urmeze și prinde orice dezvoltare. Scopul principal pentru această teză este de a prezenta o nouă cerere de extragere automată eveniment multilingv de la text (OraMEE), o astfel de chiar fiind, de exemplu, cele de mai sus menționat unul.

O sarcină și între timp cererea principală de OraMEE detectează orice eveniment legat de conflicte de la un corpus multilingv numit EventCorp și apoi depozitarea aceste evenimente într-o tabelele bazei de date (OraMEE schemă). Mai mult decât atât, abordare OraMEE este o tehnică nouă pentru extragerea unui anumit eveniment bazat pe un domeniu predefinit "arabii de primăvară". Ca orice sistem de informații de extracție (IE), OraMEE are nevoie de un corpus adnotat corespunzător pentru a instrui și evalua fiecare componente OraMEE și sub-proces. Mai mult decât atât, corpusul selectat trebuie să cuprindă, în principal de evenimente.

Datorită acestui fapt, a fost foarte dificil de a găsi un corpus propriu care conține Primăvara arabă evenimente de zi cu zi, a existat o nevoie de a dezvolta și de a construi o astfel de corpus. Prin urmare, și în scopul de a atinge obiectivul nostru, care este un OraMEE, am colectat unele dintre aceste evenimente în ceea ce am numit corpus EventCorp și am prezentat rezultatele într-un document (Alasfour & Trausan-Matu, 2013). Mai mult decât atât, acesta a fost necesar pentru EventCorp să fie aliniate (engleză, arabă) și adnotat (cu Entități Numit), prin urmare, o abordare 3D a aliniat a fost dezvoltat cu scopul de a EventCorp Aliniere (Alasfour & Trausan-Matu, 2014).