Concept Extraction

Postoje brojni pristupi za ekstrakciju informacija (*Information Extraction*) iz teksta napisanog na prirodnom jeziku, neki se baziraju na istraživanju teksta (*text minig*) i statističkom pristupu ili drugi, koji su zasnovani na principima računarske lingvistike i obrade prirodnog jezika - NLP. Prednosti statističkog pristupa su prvenstveno veća preciznost sa povećanjem korpusa za obradu, jednostavnija realizacija i duža istorija razvoja i upotrebe. S druge strane, ukoliko je potrebno obraditi kraći tekst (pitanje) kao što je to slučaj kod SIPP sistema ovakav pristup neće davati najbolje rezultate, s obzirom na mali broj reči koje se mogu analizirati. Iz tog razloga za kraće tekstove može se upotrebiti neka od tehnika računarske lingvistike. Međutim, alati za automatsku ekstrakciju informacija u principu mogu biti nedovoljno precizni i izostaviti informacije vredne za dalju analizu i izvršavanje. Takođe, kratka pitanja često mogu biti dvosmislena. Stoga, najefikasnije rešenje može se naći u vidu interaktivnog korisničkog interfejsa koji omogućava komunikaciju između modula za *obradu pitanja* i korisnika koji je to pitanje postavio. Ovaj pristup kombinuje potpuno automatsku obradu teksta i ručnu korekciju rezultata, pružajući korisniku mogućnost povećanja tačnosti izlaza. S druge strane, automatska obrada može proizvesti više rezultata koji bi obično bili zaboravljeni.

Imajući u vidu sve prethodno navedeno, kao način za predstavu dobijenih rezultata odabran je pristup u kome su otkriveni koncepti vizuelno predstavljeni u formi oblaka koncepata (*TagCloud* vizuelizacija) [**TAG07, TAG08**]. Jedna od prednosti ovog pristupa je ta da "što je koncept značajniji, ima dodeljenu veću veličinu fonta", što obezbeđuje intuitivnu predstavu specifičnih odnosa između koncepata, kao i njihovog značaja u pitanju. Takođe, generisani oblak koncepata predstavlja skup informacija koje opisuju postavljeno pitanje. Koncepti zajedno sa svojim težinama u ovom skupu čine specifičan kontekst u širem smislu koji predstavlja otisak (*fingerprint*) obrađenog teksta. Ovaj otisak je specifičan za svako pitanje i sličan je kod pitanja sa istom tematikom i istim značenjem. Na kraju, uz pomoć ovog otiska može se otkriti specifična relacija između pitanja, kao i odnos između pitanja i tema kojih se to pitanje dotiče.

### Upotrebljeni alati

Antelope (*advanced natural language object-oriented processing enviroment*) [Proxem] je skup bilioteka (alata) za obradu prirodnog jezika pod .NET okruženjem. On koristi prošireni leksikon koji nadgrađuje osnovni WordNet leksikon boljom konceptualizacijom, a takođe sadrži semantički parser i integriše formalne ontologije višeg nivoa. Povrh toga, Antelope sadrži alate za prepoznavanje imenskih entiteta (*named entity recognition*), detekciju konteksta, vremenskog perioda i lokacije, izvlačenje koreferenci, kao i mogućnost razlučivanja smisla reči (*word sense disambiguation*).

ConceptNet [CNet] je semantička mreža zasnovana na konceptima međusobno povezanih relacijama. Relacije su zasnovane na uobičajenim ljudskim saznanjima, kojih ima preko dvadeset vrsta, kao što su „služi za“, „je napravljeno od“, itd. Ovde je bitno ukazati na važan element ove semantičke mreže, koji je posebno značajan prilikom poređenja koncepata. Naime, nova saznanja su uneta u bazu od strane različith korisnika, mahom širom sveta, bez postojanja bilo kakvog ograničenja u pogledu unosa. Stoga, prikupljene podaci čine mrežu polustrukturiranih fragmenata prirodnog jezika (reči, imenskih i glagolskih grupa reči i njihovih relacija) koji predstavljaju koncepte. Takođe, za razliku od leksikona kao što je WordNet, ConceptNet u strukturi svoje baze znanja sadrži pojedine dvosmislenosti i nepreciznosti. Ovakve logičke nekorektnosti su neminovne s obzirom na prirodu samog jezika. Sa druge strane, na ovaj način ConceptNet je optimizovan za pronalaženje srodnih koncepata dok za razlučivanje dvosmislenih reči koristi pomoć leksikona. ConceptNet nudi alate za pronalaženje koncepata, povezivanje odnosno pronalaženje veza između koncepata, pronalaženje srodnih koncepata i pronalaženje teme dokumenata, kao i leksičko, sintaksno i semantičko parsiranje teksta upotrebom biblioteke za obradu teksta MontyLingua [Liu04]. Sama baza je višejezična, dok za svrhe ovog istraživanja, upotrebljen je njen deo na engleskom jeziku koji, između ostalog, ima i daleko najviše unetih koncepata i relacija.

Ceo podsitem je realizovan pomoću Microsoft .NET platforme pri čemu za izradu korisničkog interfejsa je upotrebljena Windows Presentation Fundation (WPF) tehnologija, a za potrebe komunikacije između klijentske i serverske strane je upotrebljena Windows Communication Fundation (WCF) tehnologija.

### Realizacija podsitema za analizu pitanja

S obzirom da se Antelope alat zasniva na znatno manjem, ali preciznijem WordNet leksikonu, on je predviđen za prepoznavanje imenovanih entiteta (npr. imena ljudi, organizacija, država, gradova), kao i manjeg broja koncepata. Takođe, u eksperimentalnoj fazi podržava i prepoznavanje konteksta što može doprineti boljem prepoznavanju koncepata. S druge strane, ConceptNet sadrži znatno bogatiju semantičku mrežu što daje bolje rezultate prilikom identifikovanja postojećih kao i povezanih koncepata, ali nema mogućnosti pronalaženja imenovanih entiteta niti prepoznavanja konteksta. Imajući ovo u vidu, kao i to da prilikom ekspanzije upita (*query expansion*) kombinovanje alata zasnovanih na ova dva jezička resursa (WordNet i ConceptNet) je dalo bolje rezultate u odnosu na njihovu pojedinačnu upotrebu [CNet&WordNet], modul za analizu pitanja je realizovan njihovom integracijom kao što je ilustrovano na slici 16.

Integracija alata Antelope i ConceptNet: ilustracija pristupa

S obzirom da oba alata, pored identifikovanog koncepta, određuju i njegovu težinu kao vrednost u opsegu (0,1], prilikom integracije upotrebljen je sledeći pristup: ukoliko je samo jedan od alata pronašao neki koncept, taj koncept ulazi sa svojom težinom u konačni skup, dok ukoliko su oba alata identifikovala isti koncept rezultujuća težina se računa kao njihov zbir. Ovom prilikom za računanje zbira upotebljena je probabilistička T konorma:   
*U (a, b) = a + b - a \* b*, koja ima za nijansu bržu konvergenciju ka 1 u odnosu na Ajnštajnovu T konormu: *U(a,b) = (a + b) / (1+ a\*b)*.

Na kraju kako bi se dobili bolji rezultati i smanjio broj pogrešno identifikovanih koncepata, preduzete su sledeće mere:

1. Empirijski je utvrđena linearna zavisnost između minimalne vrednosti težine identifikovanog koncepta (*threshold*) i dužine pitanja, tj. broja reči u pitanju:

Ova vrednost predstavlja granicu ispod koje se koncept izbacuje iz liste ukoliko je njegova dodeljena vrednost niža, čime se smanjuje broj pogrešno identifikovanih koncepata. Kod identifikacije konteksta nije pronađena ovakva zavisnost, pa je određena konstantna minimalna granična vrednost.

1. Pod pretpostavkom da je iz opširnog teksta teže doći do zaključka kojom se tematikom bavi, nego iz kraćeg teksta sastavljenog samo od najbitnijih pojmova, uvedena je sledeća izmena prilikom identifikacije konteksta: po pronalasku bitnih koncepata iz ConceptNet-a i svih informacija dobijenih od Antelope alata, lista ovih dobijenih informacija se ponovo prosleđuje Antelope alatu kako bi pronašao kontekst, ovog puta sa većom preciznošću.
2. Na kraju, iz skupa identifikovanih koncepata odstranjeni su oni koji predstavljaju stop reči (*stop words filtering*) – reči sa niskim informacionim sadržajem. U slučaju da ipak neke od ovih reči imaju informativni značaj korisnik ih može naknadno uneti u oblak koncepata.

GOOGLE TRANSLATE

There are many approaches to information extraction ( Information Extraction ) from a text written in natural language , based on a study of text (text minig ) and statistical approach or the other , based on the principles of Computational Linguistics and Natural Language Processing - NLP . The advantages of statistical approaches are primarily higher precision with increased corpus processing , easier to implement and longer history of development and use. On the other hand, if it is necessary to process a short text ( question ), as is the case with SIPP systems it will give the best results , given the small number of words that can be analyzed . For this reason, the short texts can be used some of the techniques of computational linguistics . However , the tools for automated information extraction in general is not sufficiently precise and leave out valuable information for further analysis and execution . Also, short questions can often be misleading . Therefore , the most efficient solution can be found in the form of an interactive user interface that enables communication between modules to handle issues and the user who asked the question . This approach combines the fully automatic processing and manual correction of the results , giving users the ability to increase accuracy of output . On the other hand , automatic processing can produce more results that would normally be forgotten.

Bearing in mind all the above , as a way to show the results chosen approach in which the detected visual concepts presented in the form of cloud concepts ( visualization TagCloud ) [ TAG07 , TAG08 ] . One advantage of this approach is that " a significant concept has assigned a larger font size ," which provides an intuitive idea of specific relationships between concepts and their importance in the matter. It also generated a cloud of concepts is a set of information that describes the question. Concepts together with their weights in this set are the specific context in a broader sense, a print ( fingerprint ) of processed text. This image is specific to each issue and similar issues with the same subject and the same meaning . Finally, with the help of this print can reveal specific relationships between issues, and the relationship between the questions and issues which this question touches .

The used tools

Antelope (advanced natural language processing object-oriented enviroment ) [ ProX ] is a library, a collection ( tools ) for natural language processing under . NET environment. He uses an extended lexicon upgrading the basic WordNet lexicon improved conceptualization and also contains semantic parser and integrates formal ontology higher level . Moreover , Antelope provides tools for named entity recognition ( named entity recognition ) , the detection of context, time period and location, drawing koreferenci , and the ability to discern meaningful words (word sense disambiguation ) .

ConceptNet [ CNet ] is a semantic network based on the concept of interconnected relationships . Relationships are based on common human knowledge, which has more than twenty species, such as " used for " , " made ​​of " and so on . It is important to point out an important element of the semantic network , which is especially important when comparing concepts. The new findings are entered into the database by the users of different , mostly around the world, without the existence of any restrictions on the entry. Therefore , the data collected form a network of semi-structured natural language fragments ( word , noun and verb groups of words and their relationships ) that represent concepts. Also, in contrast to the lexicon such as WordNet , ConceptNet in the structure of their knowledge base contains some ambiguities and inaccuracies. These logical incorrectness are inevitable given the nature of language itself . On the other hand, in this way ConceptNet is optimized for finding related concepts while the resolution of ambiguous words using the help of the lexicon . ConceptNet provides tools for finding concepts , networking and finding connections between concepts, find related concepts and finding topics of documents, as well as lexical , syntactic and semantic parsing of text using a word processing library MontyLingua [ Liu04 ] . The base itself is multilingual , and for purposes of this study , part of it was used in the English language , which, among other things, have by far the most entered the concepts and relations.

Whole Subsystem is implemented using the Microsoft . NET platform while making the user interface is used Windows Presentation Fundation (WPF ) technology , and the need for communication between client and server is used by Windows Communication Fundation (WCF ) technology .

Implementation of sub-system for the analysis of issues

Since the Antelope tool is based on a much smaller but more precise WordNet lexicon , it is designed to recognize named entities ( eg, names of people , organizations, states, cities) , as well as a small number of concepts. Also, experimental support and recognition of the context that may contribute to better identification of concepts. On the other hand, ConceptNet contains much richer semantic network which provides better results in identifying existing and related concepts , but there is no possibility of finding a named entity recognition or context. With this in mind , and that during the expansion of query ( query expansion ) combining tools based on these two linguistic resources ( WordNet and ConceptNet ) gave better results than their individual use [ CNet & WordNet ] , a module for analysis of the issue was implemented their integration as illustrated in Figure 16

Integration of tools and Antelope ConceptNet : illustration approach

Since both tools , in addition to the identified concepts, and determine the weight of a value in the range ( 0,1] , is used to integrate the following approach : If only one of the tools found some concept, this concept has with her weight in the final set while if both tools identified the same concept resulting weight is calculated as the sum . use this opportunity to compute the sum of the probabilistic upotebljena T konorma :

In (a, b ) = a + b - a \* b , which has a slightly faster convergence to 1 compared to Einstein konormu T : U (a , b ) = (a + b ) / (1 + a \* b ) .

Finally in order to obtain better results and reduce the number of incorrectly identified concepts taken the following measures :

Empirically, it is no linear relationship between the minimum weight value of an identified concept ( threshold ) and a length issue, ie . number of words in the question:

minimum weight of concepts = 8.75 % + number of words in the text \* 0.125 %

This value represents the limit below which the concept is removed from the list if its assigned value is lower , thus reducing the number of incorrectly identified concepts. In the context of identification found this dependence , and certain constant minimum limit .

Assuming that the extensive text harder to come to the conclusion that the subject matter does, but from a short text consisting only of the most important concepts introduced in the next amendment of the identification of context : the discovery of the important concepts in ConceptNet , and all information obtained from Antelope tools list of the information obtained is again forwarded Antelope tool to find context , this time with greater precision.

Finally, from the set of identified concepts removed are those that are stop words (stop words filtering ) - words with low information content . If however, some of these words have significance information user can subsequently entered into the cloud concepts.