The Polish interface for Linked Open Data

Aleksander Pohl

Computational Linguistics Department,
Jagiellonian University, Cracow, Poland
aleksander.pohl@uj.edu.pl
http://klon.wzks.uj.edu.pl/cycdemo

Abstract. This paper describes an application which aims at producing Polish descriptions for the data available as Linked Open Data, the MusicBrainz knowledge base contents in particular.

1 Introduction

It is trivial to say that the natural language is the most *natural* way of conveying information for people. No matter how many formal, unambiguous languages given person knows, when it comes to quick transfer of semantic content, natural language is always the best option. This fact is reflected in the recommendations for the user interface designers. For example one of the Nielsen's [2] usability heuristics[1] is: *The system should speak the users' language, with words, phrases and concepts familiar to the user, rather than system-oriented terms.*

This observation stays in a contrast to the fact, that Linked (Open) Data[2], which seems to be the most visible part of the Semantic Web, is usually presented in a form reflecting the structure of RDF triples[3]. And even if the content is available in a more human-readable format[4], it still looks table-like and in most cases it is only available in English.

As a result, it is hard to imagine, that a person would prefer the strict and organized information from DBpedia (who reads abstracts in several languages?) over the same information found in Wikipedia. Obviously this is not an argument against the Semantic Web, or Linked (Open) Data in particular, this is only an observation, which shows, that the way the semantic data is presented to the (non-English speaking) end users could be improved.

2 The Idea

The general idea of our approach is as follows: provide an Internet service, which allows for translation of the data available as RDF triples into natural language

[1] http://www.useit.com/papers/heuristic/heuristic_list.html

[2] For reference see: http://linkeddata.org

[3] See examples in DBpedia: http://dbpedia.org/page/Cher and Citeseer http://citeseer.rkbexplorer.com/description/resource-CS107764.

[4] See examples in OpenCalais http://d.opencalais.com/er/company/ralg-tr1r/9e3f6c34-aa6b-3a3b-b221-a07aa7933633.html and MusicBrainz http://musicbrainz.org/artist/bfcc6d75-a6a5-4bc6-8282-47aec8531818.html

descriptions. In the basic scenario, the system should be able to translate subject-predicate-object structures, such as `http://example.com/John foaf:nick ,,Johnny1''`[5] into „John's nickname is «Johnny1»". In more sophisticated scenarios, it should allow for embedding links, pictures, videos, etc. thus should be able to translate predicates such as `foaf:homepage` or `foaf:img`. And in the most advanced scenario, it should be able to produce descriptions for complex structures, such as series of events, assuring that the chronology is correct, the mentions of people and places are not replicated too often, and so on.

The value added by such a system comes form the fact, that whenever the developer of some Linked (Open) Data base uses an ontology which is known to the system, he doesn't have to write the RDF-to-NL module, since, if the meaning of the predicates used is well-defined, their natural language paraphrases should stay intact. In fact – this solution is not bound to Linked (Open) Data. With some modifications it might be used in any application with natural language (probably multilingual) interface, provided that its data structures were mapped to the above-mentioned ontology.

3 The Problems

In the case of the basic scenario, the paraphrase based on simple templates connected with particular predicates and filled with the labels of the resources or the actual values (like numbers) seems to be easily achievable. Similar approach is used in many applications with multilingual interface – when the application is localized, the templates are translated and they are filled with the data independently of the language (e.g. in Gmail there is a message at the bottom of the page: „Obecnie używasz 8 MB (0%) z 7500 MB." – „You are currently using 8MB (0%) of your 7500MB.").

In fact, this solution is not as good, as it seems, at least for inflectional languages in general and Polish in particular. The first and most problematic issue is the necessity to accommodate the gender of the subject with the verb. Thus even for the simplest sentence to be fully sound, the information required is not directly present in the RDF triple. If we consider the fact, available in the MusicBrainz knowledge base, that Cher was born on the 20^{th} of May 1946, we have to know that Cher is a women, to properly construct the sentence „Cher *urodziła* się 20 maja 1946 roku". The same fact about Michael Jackson is paraphrased as „Michael Jackson *urodził* się 29 sierpnia 1958 roku". So the template has to be adjusted with respect to the gender of the subject, but MusicBrainz doesn't contain the necessary information.

Another problem tightly connected with inflectional nature of Polish, is the inflection of numerals. In English, when some application is localized, usually there is only room for two word forms: singular and plural and the inflectional scheme is trivial (e.g. „There is one *track* on the CD/There are several *tracks* on the CD."). But in Polish numerals influence the case of the subordinate nominal

[5] See `http://xmlns.com/foaf/spec/` for the definitions of the properties.

phrase and the gender of the nominal phrase influences the form of the numeral as well. The (partial) scheme is as follows: „1(jedn*o*) krzesł*o*" (one chair) – „2(dw*a*) krzesł*a*" – „5(pięć) krzes*eł*" – „12(dwanaście) krzes*eł*" – „22(dwadzieścia dw*a*) krzesł*a*" – „25(dwadzieścia pięć) krzes*eł*"; „1(jedn*a*) ścieżk*a*" (one track) – ... The scheme is further complicated, when the numeral phrase is an argument of a verb, like in the sentence „Przysłuchiwałem się 1(jedn*ej*) ścież*ce*" (I listened to 1 track).

4 The Partial Solution

We argue that the most reasonable solution for the above mentioned problems (as well others connected with RDF-to-NL translation) is the implementation of the logic-to-NL translation system for a general ontology. Such an ontology should distinguish men from women (distinction not always present in domain-specific ontologies, like Music Ontology[6]) and other things. It should also contain many other concepts and relations, for the logic-to-NL system to be at least partially complete. Being a part of Linked Open Data is also necessary. Such a system would be a firm base for the RDF-to-NL translation application.

We think that the Cyc ontology [1] is the best candidate. First of all – it has a logic-to-English translation module, so providing the English paraphrases for RDF triples should be easy. It is a general ontology, with very broad coverage (in terms of concepts – several hundreds of thousands and relations – more than 20 thousands) and provides a Semantic Web end-point[7]. It is also well connected with DBpedia and some other elements of Linked Open Data.

Still Cyc contains only the English lexicon, that is the mapping between concepts and English words. In our previous research we created an algorithm for mapping of Cyc concepts to Polish one-segment expressions [4]. This research is further carried out and the results will be presented on the International Multiconference on Computer Science and Information Technology in the Computational Linguistic – Applications track[8]. The important fact about the constructed Polish lexicon is that, it is based on the Polish inflectional dictionary described in [3]. This means that the information necessary for accurate paraphrases accommodating common nouns and verbs as well as numerals and common nouns will be available in it.

Still the inflectional dictionary doesn't contain most of the proper names which are so common in knowledge bases and as a result the accommodation of subject and verb has to be carried out differently[9]. This is the case where the Semantic Web plays its part – the information which is so easily available for people, namely the gender of a person, might be deduced by the system from Linked Open Data and the Cyc ontology. First of all the concepts representing men and

[6] `http://musicontology.com/` – the ontology used in MusicBrainz

[7] `http://sw.opencyc.org/`

[8] `http://www.imcsit.org/pg/358/281`

[9] The inflection of unknown proper names is not covered in this research. Thus the paraphrase is not accurate if the proper name occurs at the argument position.

women should be identified in Cyc (it is #$MaleHuman and #$FemaleHuman respectively). Then the categories of the object should be looked up in the source knowledge base. If any of it is a specialization[10] of one of the above mentioned classes, the gender could be determined. If not, the same procedure should be applied for the knowledge bases containing synonyms of the object in question. The procedure would stop if any knowledge base allowed for determining the gender or certain threshold (timeout, number of visited knowledge bases, etc.) were reached.

For example Cher, whose MusicBrainz address is `http://dbtune.org/musicbrainz/resource/artist/bfcc6d75a6a5-4bc6-8282-47aec8531818`, is linked to the DBpedia resource `http://dbpedia.org/page/Cher`, where one of her OpenCyc types is `http://sw.opencyc.org/2008/06/10/concept/Mx4rvVjW5ZwpEbGdrcN5Y29ycA` (female person) which directly indicates, that she is a *woman*.

5 The Application

An application was build which creates Polish paraphrases for portion of the knowledge available in the MusicBrainz knowledge base. It is available under the URL: `http://klon.wzks.uj.edu.pl/cycdemo` and is integrated with the tool used for mapping Cyc symbols to Polish words and expressions. The actual functionality is available when the user clicks the „search" (szukaj) button and selects the „Sparql" engine. When he enters the name of an artist or an album (case sensitive), the resources found in the base are presented[11]. If he clicks the white button on the right of the resource, he will see the table with properties describing the resource. The information is presented systematically, but it is not easy to understand. If the user clicks the yellow button, he will see the Polish paraphrase of the data. Not all the data which is available in the base is presented, but the text is much more appealing and intelligible.

References

1. Lenat, D.B.: CYC: A large-scale investment in knowledge infrastructure. Communications of the ACM 38(11), 33–38 (1995)
2. Nielsen, J.: Usability Inspection Methods, chap. Heuristic Evaluation, pp. 25–62. John Wiley & Sons (1994)
3. Pisarek, P.: Słowniki komputerowe i automatyczna ekstrakcja informacji z tekstu, chap. Słownik fleksyjny, pp. 37–68. Uczelniane Wydawnictwo Naukowo-Dydaktyczne AGH (2009)
4. Pohl, A.: Automatic Construction of the Polish Nominal Lexicon for the OpenCyc Ontology. In: Recent Advances in Intelligent Information Systems (2009)

[10] The equivalence of the classes in given knwoledge base and Cyc might be established via the `http://sameas.org` service.

[11] The query takes much time, since it consists of many sub-queries to the Semantic Web end-points.