Integration of the Thesaurus for the Social Sciences (TheSoz) in an Information Extraction System

Thierry Declerck
DFKI GmbH, LT-Lab
Stuhsatzenhausweg, 3
D-66123 Saarbrücken, Germany
declerck@dfki.de

Abstract

We present current work dealing with the integration of a multilingual thesaurus for social sciences in a NLP framework for supporting Knowledge-Driven Information Extraction in the field of social sciences. We describe the various steps that lead to a running IE system: lexicalization of the labels of the thesaurus and semi-automatic generation of domain specific IE grammars, with their subsequent implementation in a finite state engine. Finally, we outline the actual field of application of the IE system: analysis of social media for recognition of relevant topics in the context of elections.

1 Introduction

Within a running research project dealing with the automatic linguistic and semantic processing of social media[1], we are working on a use case concerned with the analysis of tweets exchanged in the context of approaching election events. Besides the detection of Named Entities (name of politicians, political parties, locations, etc.) and associated opinions, we are also interested in identifying and classifying the topics people are addressing in their messages.

There are for sure topics that are very particular to a specific election, but there are also more generic and recurrent topics, some of them being of special interest to social scientists. In order to be able to detect such topics in various types of text, we have been searching for knowledge sources in the field of social and political sciences that can be used for the corresponding (both manual and automatic) semantic annotation of text. Our best candidate is for the time being the Thesaurus for the Social Sciences (TheSoz), developed by the GESIS institute at the Leibniz Institute for the Social Sciences[2]. This resource is available in the SKOS format[3], and therefore adapted to the Linked Data framework[4]. In this short paper we present first in some details the thesaurus, before describing the steps that allow us to integrate the (multilingual) language data it includes into a NLP tools suite, for the goal of supporting Knowledge-Driven analysis of texts in the field of social sciences, with a focus on micro-blogs.

2 The Thesaurus for the Social Sciences (TheSoz)

The thesaurus for social sciences is a knowledge source under continuous development (we are currently using version 0.92). The list of keywords used in TheSoz contains about 12,000 entries, of which more than 8,000 are descriptors (or "authorized keywords").

It is encoded in RDF and SKOS. While the main conceptual elements of the thesaurus are encoded in the core syntax of SKOS, the resource makes also use of the SKOS-XL properties[5] for including labels containing natural language expressions (authorized keywords, which act as domain terms) that are attached to the conceptual elements., using the "prefLabel" and "altLabel" annotation properties, allowing thus to describe main terms and their variants. The natural language expressions corresponding to the labels are encoding using the SKOS-XL annotation property "literalForm".

[1] The TrendMiner project, www.trendminer-project.eu, co-funded by the European Commission with Grant No. 287863.

[2] http://www.gesis.org/en/services/research/thesauri-und-klassifikationen/social-science-thesaurus/

[3] See http://www.w3.org/TR/skos-primer/ for a concise introduction to SKOS.

[4] http://linkeddata.org/

[5] See http://www.w3.org/TR/skos-reference/skos-xl.html

Proceedings of the 7th Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities, pages 90–95,
Sofia, Bulgaria, August 8 2013.

In order to give a (human readable) idea of the content of the thesaurus[6], we extracted with a Perl script the main elements from the SKOS code and present those in a tabular fashion, an example of which is given below, displaying also the terms in the languages covered by TheSoz (English, French and German):

concept id "10034303"

term "10034303"

- prefLabel id "10034303"
- lang=de "Abbrecher"
- lang=en "drop-out"
- lang=fr "drop-out"
- altLabel id "10034307"
- lang=de "Studienabbrecher"
- lang=en "university drop-out"
- lang=fr "étudiant qui abandonne ses études"

notation „3.2.00"

- lang=de „Schule und Beruf (berufliche Qualifikationselemente im Bereich der schulischen Ausbildung)"
- lang=en "School and Occupation (Elements of Occupational Qualification in School Education)"
- lang=fr « École et profession (éléments de qualification professionnelle dans le domaine de l'enseignement scolaire) »

broader notation „3.2"

- lang=de „Beruf und Qualifikation"
- lang=en „Occupation and Qualification"
- lang=fr « profession et qualification »

broader notation „3"

- lang=de „Interdiszipinäre Anwendungsbereiche der Sozialwissenschaften"
- lang=en "Interdisciplinary Application Areas of Social Sciences"
- lang=fr « domaines interdisciplinaires d'application des sciences sociales »

In the example above the reader can see how the English preferred label "drop-out" is associated with the concept "School and Occupation", which is itself a subclass of the concept "Occupation and Qualification", classified itself as a field of the broader concept "Interdisciplinary Application Areas of Social Sciences". All the language material contained in the labels or used for naming the "notations" can be re-used for detecting and semantically annotating the related topics in running texts.

3 TheSoz as Linked Data

The encoding of TheSoz in SKOS is an important asset, since it allows linking the data to other knowledge sources, like for example DBpedia[7] in the Linked Data framework, and so to complement information contained in TheSoz, which remains at the terminological level, and is thus not giving detailed information about the included multilingual terms for the described concepts and the relations between those.

So for example TheSoz mentions the main political parties in Germany, Austria and other countries, but not their actual leader, their actual role (in the government or in the opposition) or weight in the current legislation period. TheSoz also lists the names of important persons, like "Merkel, A." or "Brandt, W.", but no biographical indication or relation to political parties or institutions are given. As such TheSoz is providing for a light-weight ontological basis, with multilingual labels, which allows detecting in text mentions of topics or entities relevant to the social scientists.

The linking of concepts and associated terms to more elaborated knowledge sources, like DBpedia, is thus necessary in order to implement a full Knowledge Driven Information Extraction (KDIE) system in the field of social sciences. So for example the TheSoz sub-term "university" in "university drop-out" can be completed by information in the DBpedia entry for "university", stating among others that "university" is *rdfs domain* of "numberOfPostgraduateStudents" and that it is a *subClassOf* "EducationalInstitution. "http://schema.org/EducationalOrganization" is given as an *equivalenceClass* of the DBpedia entry for "EducationalInstitution". From the schema.org entry we can make use of additional relations associated to "EducationalInstitution", like for example a relation to more specific types, such as "CollegeOrUniversity", "ElementarySchool", "HighSchool", "MiddleSchool", "Preschool", "School". We can this way expand the terminological base of TheSoz by accessing the labels of the classes and concepts of other knowledge sources referred to by explicit semantic relations like *owl*:*equivalentClass*, *owl*:*sameAs* or *skos*:*exactMatch*.

As the reader can see from the name of the mentioned ontology classes above, natural language expressions associated to elements of knowledge sources can have different surface forms as the one we saw in the examples of "literalForms" of TheSoz. Beyond the utilization of the annotation properties, such as rdfs:label,

[6] Online visualizations and access are available at http://lod.gesis.org/thesoz/

[7] See http://dbpedia.org/About. And in fact, 5024 TheSoz concepts are linked to DBpedia via SKOS "exact matches".

skosxl:prefLabel" or skosxl:literalForm, dedicated to ease the understanding by human users, several other syntax elements of knowledge representation systems, such as the RDF URI references, like rdf:ID, rdf:about, or rdf:resource, may contain instead of numerical codes natural language expressions, often using the CamelCase notation. Fu et al. (2012) describes NLP tasks and applications using natural language expressions contained in such RDF URI references. In our work, we focus on natural language expressions contained in the annotation properties rdfs:label, sxkos:label (skosxl:prefLabel and others) and skosxl:literalForm, which typically include textual material to be consumed by human readers, and which can be normally directly processed by NLP tools, without requiring prior transformation processes of the textual material.

4 Integration of TheSoz in a NLP Framework

Before applying the (possibly extended) terminological material of TheSoz for supporting the semantic annotation of running texts, it has to be submitted to pre-processing steps, in order to ensure as a minimum a possible matching to morpho-syntactic variations of (elements of) the terms that are to be expected in external text. For this, we need to lexicalize the labels of the thesaurus, transforming the terms to linguistic data that can be used for matching linguistically processed text. A first sketch of this approach has been described in (Declerck & Lendvai, 2010) and a more elaborated methodology, encoding the linguistic data in RDF is presented in (McCrae et al, 2012).

And for ensuring a linking of linguistic data in text to the conceptual elements of the thesaurus (or other knowledge sources), the development of an information extraction grammar is needed. We present in section 3.2 below an automatized approach for this.

For both steps we are using the NooJ platform[8], whose finite states engine supports the flexible implementation of lexicons, morphological, syntactic and semantic grammars.

4.1 Lexicalization

The lexicalization step consists in submitting all the language material included in the knowledge source to a lexical and a syntactic analyzer, which in our case are lexicons and grammars implemented in NooJ.

The results of such a processing can be encoded in the lexicon-ontology model *lemon* (McCrae et al, 2012), which declaratively represents textual and linguistic information of ontologies as additional RDF resource linked to the original concepts associated to the labels. The *lemon* model decomposes multi-word expressions to individual words and represents the results in a phrase structure, which can be shared by multiple lexical entries. Furthermore, dependency relations between decomposed phrase constituents can be modeled. A simplified example of the *lemon* representation of the NooJ parsed term "university drop-out" is shown below:

```
:university_drop-out [lemon:writtenRep "univer-
sity drop-out"@en]
lemon:sense [lemon:reference ontol-
ogy:TheSoz10034307];
lemon:decomposition ( :university_comp
:drop-out_comp ) ;
lemon:phraseRoot [ lemon:constituent :NP ;
lemon:edge [lemon:constituent :NP ;
lemon:edge [lemon:constituent :NN ;
lemon:leaf university_comp ] ;
lemon:edge [lemon:constituent :NN ;
lemon:leaf drop-out_comp ] ];
].
```

For the sake of simplicity we do not display the *lemon* representation of additional analysis provided by NooJ (for example the one, which is decomposing "drop-out" in two lemmas). It is enough to mention that *lemon* also supports the representation of preferred and alternative labels. This is important if one wants to consider all possible (linguistically annotated) term variants for improving the matching of TheSoz terms to terminological variants in text, going thus beyond the matching of terms to purely morpho-syntactic variations. So for example, in TheSoz "drop-out" is the prefLabel, while "university drop-out" is marked as altLabel of the same concept. Such term variants can also be "imported" in our lexicalization step from other source. Or one can import additional lexical material, so for example the corresponding WordNet synonyms or glosses. In the next future we also plan to "tap" the BabelNet[9] resource, which is providing links to WordNet, Wikipedia and DBpedia (and more is planed), for extending the terminological

[8] http://www.nooj4nlp.net/pages/nooj.html

[9] See http://lcl.uniroma1.it/babelnet/ or (Navigli & Ponzetto, 2012).

base of the (lexicalized) TheSoz labels, also with terms in languages not covered by TheSoz for now.

4.2 Automatic Generation of Domain specific IE grammars

On the basis of the lexicalization step described in section 3.1, we wrote a Perl program that generates IE grammars in the NooJ finite state engine. This procedure is done in 5 steps.

1) Using the Term ID of TheSoz as names for NooJ recognition rules.
   term10034307 =
2) Using the corresponding lexicalised labels as the expressions to be recognized by the NooJ rule (abstract representation):
   term10034307 = [lemma=„university“ cat=„N“] [lemma=„drop-out“ cat=„N“] ;
3) Adding possible term variants to the rule) [10]:
   term10034307 = ([lemma=„university“ cat=„N“] [lemma=„drop-out“ cat=„N“] | :var10034307) ;

   var10034307 = [lemma=„university“ cat=„N“] [lemma=„drop“ cat=„V“] [lemma=„out“ cat=„P“] ;
4) Linking the linguistically annotated prefLabel and the altLabel(s) to the corresponding Concept ID, as the basis of the semantic organization of the lexical material in NooJ:
   concept10034303 = (term10034303 | term10034307) ;
5) Defining the annotation generation procedure of the NooJ rules: Successful application of the rule *concept10034303* can generate the following annotation:
   CLASS= TheSoz_ID=“10034303” altLabel_ID=“10034307“ altLabel =“universtiy drop-out@en“ SuperClass=TheSoz_ID_3.2 SuperClassLabel = „Occupation and Qualification“ altLabel_Translation = „Studienabbrecher@de“ etc.[11])

This procedure has been fully implemented, using Perl scripts. The addition of term variants (in red color in the example above, point 3) can be done manually or automatically. We are also currently adding information about the context of such terms to be expected in running texts, like for example the agent of the event “drop-out”, and further modifications, like date, location and reasons.

At the moment we are able to semantically disambiguate in text for example the two senses of the TheSoz term “drop-out”: one in the sense of “university drop-out” and the one in the sense of “resignation from occupation”. The generated NooJ grammars are currently being tested for a use case dealing with the elections in Austria.

5 Use Case

Our actual focus is the elections in Austria. Our aim is to detect which topics are of have been discussed in the social media, and how this relates to election results obtained by candidates and parties.

As such we cannot report yet on evaluation results, both at the technological and usability level, since an evaluation study is still to be performed. We will be using collection of polls for measuring the accuracy of the detection of topics and the related popularity of parties/politicians detected in social media.

The use case partner involved in the project has been designing an annotation schema and is performing a semi-automatic annotation of selected tweets and blogs, which we will use as gold standard.

A fully operational system is expected to work for the national elections in Austria to be held on the 28th September of 2013.

6 Future Work

Besides the evaluation work sketched in the former section, the next steps in our work will consist in aggregating information from other

[10] In this simplified example we do just include as a term variant the decomposition of the noun “drop-out” in two lemmas, extending thus the lexical coverage of the original label. The final rule (not displayed here for the sake of simplicity) is also stating that the sub-term “university” doesn’t have to immediately precede the sub-term “drop”, accounting thus also for alternative word order.

[11] An example text is: “Mar 29, 2012 – Record numbers of students quit *university courses* last year as the higher education *drop-out* rate soared above 30000 for the first time…”

knowledge source, not only from DBpedia but also from a recently developed political ontology, which has been designed in the context of our project.

We have also already conducted experiments in relating the linguistically annotated terms of TheSoz with terms available in other thesauri, like for example GEMET[12]. As GEMET is containing labels in 33 languages, this linking will allow us to find more multilingual equivalents of terms in TheSoz, at least for the concepts of TheSoz that can be associated with concepts in GEMET.

Another line of investigation will consist in adapting the work on correcting and complementing the labels used in TheSoz, following the reports described in (Declerck & Gromann, 2012), where correcting and completive patterns have been applied to the labels of multilingual taxonomies dealing with the description of industry activity fields of companies listed in various stock exchanges. Improving the terminological quality of labels seems to be a good strategy for improving knowledge-driven information extraction.

Following the approaches to cross-lingual harmonization of taxonomy labels described in (Declerck & Gromann, 2012; Gromann & Declerck, 2013), we notice that in many multilingual knowledge sources (Thesauri, Taxonomies or Ontologies), the content of multilingual labels is not parallelized. In one of our example within the TheSoz, displayed in Section 2, we had the following concept with the labels in three languages:

```
term "10034303"
   concept  id "10034303"
      …
      altLabel id "10034307"
      altLabel de "Studienabbrecher"
      altLabel en "university drop-out"
      altLabel fr "étudiant qui abandonne ses études"
      ….
```

As the reader can see, only the French label is containing explicitly the fact the entity "performing" the drop-out is a student. Although the super-classes make clear that "university drop-out" is in the field of "School and Occupation", none of the metadata or labels, other as the French "altLabel" is mentioning that a student is involved in this field. The German label can lead to the reading that a person is involved, if adequate lexical semantics resources are used. The English label does not mention at all that an agent is involved: it just names the event. The French and German labels are about abandoning "studies" while the English label is about abandoning "university".

As suggested by Gromann & Declerck (2013), we can add (either manually or by automated process) to the English alternative labels the translations of the French label (in this particular case, the one with the richest contextual information), like "a student, who is dropping out his studies". This is important since it improves the matching of the concepts of TheSoz to running texts.

7 Conclusion

We have described actual work in integrating multilingual knowledge sources in the field of social sciences into a NLP task, consisting in identifying relevant topics of discussion in social media. As it is still too early to report on results (due to the internal calendar of the project), we could only present for the time being the current state of implementation, which consisted in first lexicalizing the labels of the knowledge source "TheSoz", freely available – in the SKOS format. On the basis of the lexicalized labels, and their relation to conceptual element of the knowledge source, we implemented an automatic generation of knowledge-driven IE grammars, which have been realized as finite state transducers in the NooJ platform. Those resulting IE grammars are to be deployed in the context of a use case dealing with the detection of topics addressed in social media on approaching elections.

Acknowledgments

The work presented in this paper has been supported by the TrendMiner project, co-funded by the European Commission with Grant No. 287863.
The author is thanking the reviewers for their very helpful comments, which led to substantial changes brought to the final version of the paper.
The author is also thanking Dagmar Gromann (Vienna University of Economics and Business). Intensive discussions with her on related topics have been heavily inspiring the work described in this paper.

[12] GEMET stands for "GEneral Multilingual Environmental Thesaurus". See also http://www.eionet.europa.eu/gemet/

References

Declerck, T., Lendvai, P. 2010. Towards a standardized linguistic annotation of the textual content of labels in Knowledge Representation Systems. In: *Proceedings of the seventh international conference on Language Resources and Evaluation*, Valetta, Malta, ELRA.

Fu, B., Brennan, R., O'Sullivan, D.: A Con_gurable Translation-Based Cross-Lingual Ontology Mapping System to Adjust Mapping Outcomes. Journal of Web Semantics, Vol. 15, pp.15_36 (2012)

Declerck, T., Gromann, D. 2012. Towards the Generation of Semantically Enriched Multilingual Components of Ontology Labels. In: *Proceedings of the 3rd Multilingual Semantic Web Workshop*.

Ell, B., Vrandecic, D., Simperl, E. 2011. Labels in the Web of Data. In Aroyo, L., Welty, C., Alani, H., Taylor, J., Bernstein, A. (eds.): *Proceedings of the 10th international conference on the semantic web* - Volume Part I (ISWC'11), Vol. Part I. Springer-Verlag, Berlin, Heidelberg, pp.162_176.

Fu, B., Brennan, R., O'Sullivan, D.: A Configurable Translation-Based Cross-Lingual Ontology Mapping System to Adjust Mapping Outcomes. Journal of Web Semantics, Vol. 15, pp.15_36 (2012)

Garcia, J., Montiel-Ponsoda, E., Cimiano, P., Gómez-Pérez, A., Buitelaar, P., Mc- Crae, J. 2012. *Challenges for the Multilingual Web of Data*. Web Semantics: Science, Services and Agents on the World Wide Web, Vol. 11, pp.63-71.

Gromann, D., Declerck, T. 2013. Cross-Lingual Correcting and Completive Patterns for Multilingual Ontology Labels. In Buitelaar, P. and Cimiano, P. (eds) *Multilingual Semantic Web*, Springer-Verlag (to appear)

McCrae, J., Aguado-de-Cea, G., Buitelaar, P., Cimiano, P., Declerck, T., Gómez-Pérez, A., Gracia, J., Hollink, L., Montiel-Ponsoda, E., Spohr, D., Wunner, T. 2012. *Interchanging lexical resources on the SemanticWeb.* Journal of Language Resources and Evaluation, pp.1_19.

Navigli, N., Ponzetto, S.P.. 2012. *BabelNet: The Automatic Construction, Evaluation and Application of a Wide-Coverage Multilingual Semantic Network.* Artificial Intelligence, 193, Elsevier, pp. 217-250.

Silberztein, Max. 2003. *NooJ manual*. Available at the WEB site http://www.nooj4nlp.net (200 pages)

Wimalasuriya, D. C., Dou, D. 2012. *Ontology-based information extraction: an introduction and a survey of current approaches.* Journal of Information Science, Vol. 36, No. 3, pp.306-_323.

Zapilko, B., Johann Schaible, Philipp Mayr, Brigitte Mathiak. 2012. *TheSoz. A SKOS Representation of the Thesaurus for the Social Sciences*. Semantic-Web Journal.