RECSA: Resource for Evaluating Cross-lingual Semantic Annotation

Achim Rettinger*, Lei Zhang*, Daša Berović**,
Danijela Merkler, Matea Srebačić**, Marko Tadić*****

* Karlsruhe Institute of Technology
76128 Karlsruhe, Germany
{rettinger, l.zhang}@kit.edu
** University of Zagreb
Trg maršala Tita 14, 10000 Zagreb, Croatia
{dasa.berovic, danijela.merkler}@ffzg.hr, {msrebaci}@unizg.hr
***Faculty of Humanities and Social Sciences, University of Zagreb
Ivana Lučića 3, 10000 Zagreb, Croatia
marko.tadic@ffzg.hr

Abstract

In recent years large repositories of structured knowledge (DBpedia, Freebase, YAGO) have become a valuable resource for language technologies, especially for the automatic aggregation of knowledge from textual data. One essential component of language technologies, which leverage such knowledge bases, is the linking of words or phrases in specific text documents with elements from the knowledge base (KB). We call this semantic annotation. In the same time, initiatives like Wikidata try to make those knowledge bases less language dependent in order to allow cross-lingual or language independent knowledge access. This poses a new challenge to semantic annotation tools which typically are language dependent and link documents in one language to a structured knowledge base grounded in the same language. Ultimately, the goal is to construct cross-lingual semantic annotation tools that can link words or phrases in one language to a structured knowledge database in any other language or to a language independent representation. To support this line of research we developed what we believe could serve as a gold standard Resource for Evaluating Cross-lingual Semantic Annotation (RECSA). We compiled a hand-annotated parallel corpus of 300 news articles in three languages with cross-lingual semantic groundings to the English Wikipedia and DBPedia. We hope that this new language resource, which is freely available, will help to establish a standard test set and methodology to comparatively evaluate cross-lingual semantic annotation technologies.

Keywords: cross-lingual semantic annotation, evaluation resource, Linked Open Data

1. Introduction

The ever-increasing quantities of semantic data on the Web pose new challenges but at the same time open up new opportunities of publishing and accessing information on the Web. Language technologies can both, benefit from and support linked data repositories. For instance, knowledge extraction from text can utilize the facts in DBpedia (Bizer et al., 2009), Freebase (Bollacker et al., 2008), or Yago (Hoffart et al., 2013) as seed knowledge for the discovery of the relevant extraction patterns in large volumes of texts (Krause et al., 2012). On the other hand those technologies can help to grow the knowledge base by automatic extraction of knowledge from text documents. From the side of language technologies, wordnets (Fellbaum, 1998) or automatic ontology population methods (Buitelaar and Cimiano, 2008) represent similar resources and techniques.

At the core of such technologies is the ability to relate words and phrases in natural language texts to existing resources in a knowledge base. If successful, a semantically annotated text document allows automatic contextualization and inference about the content of the document. Obviously, this task is highly language dependent, both on the side of the text document and the specific language interface to the knowledge base. In order to achieve the goal that speakers of all languages have access to the same information, there is an impending need for systems that can help in overcoming language barriers by facilitating multilingual and cross-lingual access to basic semantic data originally produced for a different culture and language.

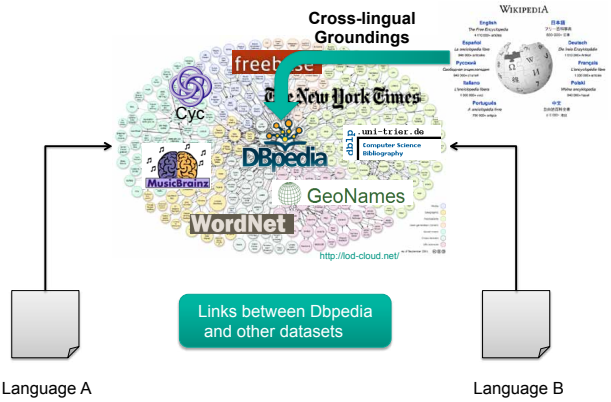

Figure 1: Cross-lingual Semantic Annotation with DBpedia

In order to connect multilingual information across languages, efforts have been made on two levels: on the one hand, machine translation systems can connect multilingual text documents to each other, and on the

other hand, multilingual KB resources have been linked across languages (e.g. through language links in Wikipedia) or have been lifted to a language independent representation (e.g. Wikidata).

By combining techniques from both levels, the ultimate goal should be to construct cross-lingual semantic annotation tools that can link words and phrases in one language to structured knowledge in any other language or to a language independent representation.

There have been extensive analyses of each of the tasks separately: a) for machine translation evaluation efforts, see for instance (Papineni et al. 2002; Han et al. 2012), b) semantic annotation evaluation efforts (McNamee et al 2009; TAC_KBP [1] in 2013 address multilingual entity linking but not cross-lingual linking) and c) Wikipedia cross-language links analysis (Melo et al. 2010; Rinser et al. 2012). However, there have not been any attempts in evaluating cross-lingual semantic annotation tools as a whole, partially because the evaluation test set for this particular task was missing.

In this paper we propose a resource that can be used as a standard test set and procedure for evaluating and benchmarking cross-lingual semantic annotation systems. Our contribution is two-fold: First, in Section 2. we introduce our hand-labeled data set with words and phrases linked to English Wikipedia articles and indirectly to DBpedia resources. Second, in Section 3. we describe how this data set can be used to evaluate cross-lingual semantic annotation techniques. In Section 4 we provide conclusion and mention possible future development.

2. Cross-lingual Semantic Annotation Evaluation Data

The RECSA resource we have developed consists of 300 news articles in three different languages (English, German and Spanish) with 100 articles in each language. The source of texts is a non-profit community of authors and translators Global Voices portal[2], that brings news reports in 30+ different languages and make their texts available under CC-BY license. All 100 articles run in parallel, thus forming a trilingual aligned parallel corpus, and therefore allowing the investigation of cross-lingual semantic annotation techniques that keep the multilingual content under control, i.e. the content is the same in all three languages. The articles were downloaded, their boiler-plates were removed and they were converted into a plain text for further processing.

	English	German	Spanish
Tokens in total	74337	79146	77476
Sentences per article	27.12	34.62	27.36

Table 1. Basic statistics on RECSA resource

[1] http://www.nist.gov/tac/2013/KBP/data.html
[2] http://www.globalvoicesonline.org

The final desired result of producing RECSA was to have a resource that will have (1) NEs detected, and classified; (2) general concepts mentioned in text also detected. Both of these "lexical groundings" are then linked to their respective Wikipedia articles. Having all this annotated in a trilingual parallel corpus, will provide opportunity to the researchers to measure the quality of their systems when they establish links between texts and Wikipedia articles in a monolingual and cross-lingual context.

English Wikipedia was selected as the most encompassing conceptual resource that is explicitly cross-linked to many languages, i.e. expressed in many languages. For this reason English Wikipedia was used as a hub conceptual space that also exhibits direct one-to-one links to DBpedia, which again, is linked with WordNet and numerous other linked data sources (see Figure 1). The lexical groundings in German and Spanish documents are also linked to English Wikipedia.

```
<sentence id="1">
    <text>An internationally renowned Iranian filmmaker, Mohsen Makhmalbaf, outrage
    <tokens>
        <token pos="Z" end="2" lemma="1" id="1.1" start="0">An</token>
        <token pos="RB" end="18" lemma="internationally" id="1.2" start="3">interna
        <token pos="JJ" end="27" lemma="renowned" id="1.3" start="19">renowned</tok
        <token pos="NP00V00" end="35" lemma="iranian" id="1.4" start="28">Iranian</
        <token pos="NN" end="45" lemma="filmmaker" id="1.5" start="36">filmmaker</
        <token pos="Fc" end="46" lemma="," id="1.6" start="45">,</token>
        <token pos="NP00SP0" end="64" lemma="mohsen_makhmalbaf" id="1.7" start="47"
        <token pos="Fc" end="65" lemma="," id="1.8" start="64">,</token>
        <token pos="VBD" end="74" lemma="outrage" id="1.9" start="66">outraged</tok
        <token pos="PRP" end="79" lemma="many" id="1.10" start="75">many</token>
        <token pos="NP00V00" end="88" lemma="iranians" id="1.11" start="80">Iranian
        <token pos="IN" end="91" lemma="by" id="1.12" start="89">by</token>
        <token pos="VBG" end="101" lemma="accept" id="1.13" start="92">accepting</
        <token pos="Z" end="104" lemma="1" id="1.14" start="102">an</token>
        <token pos="NN" end="115" lemma="invitation" id="1.15" start="105">invitat
        <token pos="TO" end="118" lemma="to" id="1.16" start="116">to</token>
        <token pos="DT" end="122" lemma="the" id="1.17" start="119">the</token>
        <token pos="NP00V00" end="146" lemma="jerusalem_film_festival" id="1.18" st
        <token pos="IN" end="149" lemma="in" id="1.19" start="147">in</token>
        <token pos="NP00G00" end="156" lemma="israel" id="1.20" start="150">Israel
        <token pos="DT" end="161" lemma="this" id="1.21" start="157">this</token>
        <token pos="NN" end="167" lemma="month" id="1.22" start="162">month</token
        <token pos="Fp" end="168" lemma="." id="1.23" start="167">.</token>
    </tokens>
</sentence>

<annotation displayName="Israel" entityId="5" weight="0.925">
    <descriptions>
        <description URL="http://en.wikipedia.org/wiki/Israel" lang="en"/>
        <description URL="http://de.wikipedia.org/wiki/Israel" lang="de"/>
        <description URL="http://es.wikipedia.org/wiki/Israel" lang="es"/>
    </descriptions>
    <mentions>
        <mention sentenceId="1" words="Israel"/>
        <mention sentenceId="5" words="Israel"/>
        <mention sentenceId="7" words="Israel"/>
        <mention sentenceId="9" words="Israel"/>
        <mention sentenceId="11" words="Israel"/>
        <mention sentenceId="12" words="Israel"/>
    </mentions>
</annotation>
```

Fig 2. Illustration of the processed text in XLike format

The first step in producing the RECSA resource was to get the parallel corpus annotated for Named Entities (NEs), including their categories (Location, Person, Organization, Other).

The second step consisted of adding additional annotation with links of NEs and general concepts to Wikipedia. After each step a manual verification and cleaning up was conducted.

The plain text documents were first processed using the XLike project[3] linguistic processing pipelines (Padró et al. 2014) for English, German and Spanish in order to receive automatic stand-off NE annotation. This annotation was then manually verified and cleaned. The pipelines consist of tokenization module, POS-tagging and lemmatization module and NERC module, each running as a separate, but mutually connected web service.

[3] http://www.xlike.org

The following step was the application of a semantic annotation method developed in the XLike project, based on a newly developed cross-lingual linked data lexica, called xLiD-Lexica[4]. We constructed xLiD-Lexica by exploiting the multilingual Wikipedia to extract the cross-lingual groundings of resources in KBs, which are also called surface forms, i.e. terms (including words and phrases) in different languages that can be used to refer to resources. Besides the extracted surface forms, we also exploit statistics of the cross-lingual groundings to measure the association strength between surface forms in different languages and the referent resources (Zhang et al. 2014). The results were added to the cleaned output from the linguistic processing pipelines and all possible general concepts (GCs) "lexical groundings" or mentions were marked. The automatic processing was targeted to receive the highest possible recall, so this step provided a noisy output with a lot of links to Wikipedia articles, many of them incorrect. This output was manually verified and cleaned to achieve the cleanest possible resource. As before with NEs, the links for detected GCs were pointing primarily to English Wikipedia articles.

	English	German	Spanish
Automatic NEs	4629	2768	1535
Automatic GCs	14847	10303	9426
Correct(ed) NEs	1008	874	1050
Correct(ed) GCs	3866	3113	3842
Correct(ed) +NEs	0	3	3
Correct(ed) +GCs	0	5	5
Correct(ed) -NEs	329	395	589
Correct(ed) -GCs	382	384	658
Correct(ed) ?NEs	32	24	37
Correct(ed) ?GCs	545	484	842

Table 2. Statistics on semantic annotations (wiki-links): NE = Named Entity; GC = General Concept

The manual verification and cleaning process consisted of checking whether the automatic processing detected mentions of real NEs and GCs and whether they exist in Wikipedia (primarily in English and secondary in German or Spanish). The rules for manual checking were designed to exhaustively cover different possibilities:

1. NE exists in en.Wikipedia:
   a. if the mention and en.Wikipedia article title are the same, insert [NE[Mention]]
   b. if the mention and en.Wikipedia article title are not the same, insert [NE[Title|Mention]]
2. GC exists in en.Wikipedia:
   a. if the mention and en.Wikipedia article title are the same, insert [GC[Mention]]
   b. if the mention and en.Wikipedia article title are not the same, insert [GC[Title|Mention]]
3. NE or GC cannot be found in en.Wikipedia (target language), but can be found in de.Wikipedia or es.Wikipedia (source language):
   a. for NE insert [+NE[Mention]] or [+NE[Title|Mention]]
   b. for GC insert [+GC[Mention]] or [+GC[Title|Mention]]
4. NE or GC cannot be found in any of Wikipedias, en, de or es (both source language and target language):
   a. for NE insert [-NE[Mention]]
   b. for GC insert [-GC[Mention]]
5. The exact NE and GC matching the mention cannot be found in Wikipedia, but the very related ones can:
   a. for NE insert [?NE[Title|Mention]]
   b. for GC insert [?GC[Title|Mention]]

The manual verification and cleaning were performed for each of 300 documents by two different human annotators. The calculated anotator agreement is shown in Table 3.

	English	German	Spanish
All types of NEs	0.981	0.833	0.991
All types of GCs	0.970	0.861	0.801

Table 3. Statistics of annotator agreement over different types of annotations

The manual annotations were automatically converted into the XLike XML stand-off annotation format that RECSA is using. The detailed description of the format is available as a public deliverable from the XLike project web site. The RECSA resource will be availabe through META-SHARE[5] under permissive license.

3. Cross-lingual Semantic Annotation Evaluation Methodology

With the availability of the RECSA resource, a standard evaluation methodology for cross-lingual semantic annotation can be conducted. Different semantic annotation systems can use RECSA for measuring their quality since all NEs and all GCs found in the documents are marked and linked explicitly to the conceptual space (Wikipedia). The cross-lingual annotation can be evaluated by the number of detected links to the same Wikipedia article by a new system, in comparison to the links existing in RECSA in any of different three languages. This way, the robustness of a method in regard of its performance for different languages can also be measured.

Since the NEs and GCs are annotated inside the text documents by using XML markup, it is straightforward for evaluators to predict all links within the `<entities>…</entities>` element and automatically compare the outputs. Basic evaluation measures can be calculated as follows:

- True Positives are the correctly detected and linked mentions.
- False Positives are the incorrectly detected or linked mentions.

[4] http://km.aifb.kit.edu/services/xlike-lexicon/

[5] http://www.meta-share.eu

- False Negatives are the incorrectly not detected and linked mentions.

Based on these values, well established measures like Precision, Recall, F-Measure or ROC can be calculated.

4. Conclusions and future work

We have presented a Resource for Evaluating Cross-lingual Semantic Annotation (RECSA). Cross-lingual semantic annotation is essential to many modern language technologies, but there was a lack of validated benchmarks to test them.

We compiled a hand-annotated parallel corpus of 300 news articles in three languages with cross-lingual "lexical groundings" to the English Wikipedia and indirectly to DBpedia. The resource can be used to evaluate the performance for every language separately, compare the performance for different languages or get the average performance across languages. Furthermore, since DBpedia is the nucleus of the Linked Data Cloud, annotations with numerous structured data sets can be evaluated, as long as they are linked to DBpedia and indirectly to Wikipedia.

We hope, that this new language resource will help to establish a standard test set and contribute to the development of methodology to comparatively evaluate cross-lingual semantic annotation techniques.

Our future work is concerned with the extensions to additional languages and additional semantic resources. Furthermore, we are considering the possibility to hand-label the semantic roles encoded by verbs and relations on top of the NEs and GCs and in that way open the possibility for evaluation of systems that detect more complex relations between entities from KB in texts.

5. Acknowledgements

The research leading to these results was supported by the European Community's Seventh Framework Programme FP7-ICT-2011-7, project XLike, grant no. 288342.

6. References

Christian Bizer, Jens Lehmann, Georgi Kobilarov, Sören Auer, Christian Becker, Richard Cyganiak, and Sebastian Hellmann. 2009. Dbpedia-a crystallization point for the web of data. *Web Semantics: Science, Services and Agents on the World Wide Web*, 7(3):154–165.

Paul Buitelaar, Philipp Cimiano. 2008. *Ontology learning and population: bridging the gap between text and knowledge*, Ios Press, Amsterdam.

Kurt Bollacker, Colin Evans, Praveen Paritosh, Tim Sturge, and Jamie Taylor. 2008. Freebase: a collaboratively created graph database for structuring human knowledge. *In Proceedings of the 2008 ACM SIGMOD international conference on Management of data*, ACM, pp 1247–1250.

Fellbaum, Christianne (ed.) 1998. *Wordnet: An electronic lexical database*, MIT Press, Cambridge MA.

Johannes Hoffart, Fabian M Suchanek, Klaus Berberich, and GerhardWeikum. 2013. Yago2: a spatially and temporally enhanced knowledge base from wikipedia. *Artificial Intelligence*, 194:28–61.

Sebastian Krause, Hong Li, Hans Uszkoreit, and Feiyu Xu. 2012. Large-scale learning of relation-extraction rules with distant supervision from the web. In *The Semantic Web–ISWC 2012*, Springer, pp 263–278.

McNamee, P. & Dang, H. T. 2009. Overview of the TAC 2009 Knowledge Base Population Track. In *Proceeding of Text Analysis Conference*.

Gerard de Melo, Gerhard Weikum. 2010. Untangling the cross-lingual link structure of Wikipedia. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, Association for Computational Linguistics

Lluís Padró, Željko Agić, Xavier Carreras, Blaž Fortuna, Esteban García-Cuesta, Zhixing Li, Tadej Štajner, Marko Tadić. 2014. Language Processing Infrastructure in the XLike Project. Proceedings of the 9th Language Resources and Evaluation Conference (LREC2014), ELRA, Reykjavik-Paris, (in this volume).

Kishore Papineni, Salim Roukos, Todd Ward, Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. Proceedings of the 40th Annual meeting of the Association for Computational Linguistics (ACL-2002), ACL, pp 311–318.

D. Rinser, D. Lange, F. Naumann. 2012. Cross-lingual entity matching and infobox alignment in Wikipedia. In *Information Systems*, Elsevier.

Lei Zhang, Färber Michael, Achim Rettinger. 2014. xLiD-Lexica: Cross-lingual Linked Data Lexica. Proceedings of the 9th Language Resources and Evaluation Conference (LREC2014), ELRA, Reykjavik-Paris, (in this volume).